**Thesis for the Degree of Doctor of Philosophy**

# Haptic Perception Modeling and Signal Generation for Textured Surfaces and Car Doors using Deep Learning

**Mudassir Ibrahim Awan**

**Department of Computer Science & Engineering**
**Graduate School**
**Kyung Hee University**
**South Korea**

**August 2025**

**Thesis for the Degree of Doctor of Philosophy**

# Haptic Perception Modeling and Signal Generation for Textured Surfaces and Car Doors using Deep Learning

**Mudassir Ibrahim Awan**

**Department of Computer Science & Engineering**
**Graduate School**
**Kyung Hee University**
**South Korea**

**August 2025**

# Haptic Perception Modeling and Signal Generation for Textured Surfaces and Car Doors using Deep Learning

by

**Mudassir Ibrahim Awan**

Supervised by

**Prof. Seokhee Jeon, Ph.D.**

Submitted to the Department of Computer Science and Engineering and the Faculty of the Graduate School of Kyung Hee University in partial fulfilment of the requirements for the degree of Doctor of Philosophy

Dissertation Committee:

**Prof. Sung-Ho Bae, Ph.D.** ( Chairman ) .....

**Prof. Seungjae Oh, Ph.D.** .....

**Prof. Jung Uk Kim, Ph.D.** .....

**Prof. Yongjae Yoo, Ph.D.** .....

**Prof. Seokhee Jeon, Ph.D.** .....

# Abstract

Human haptic perception is fundamentally multidimensional, shaped by both tactile cues such as texture and vibration, and kinesthetic signals such as motion, force, and stiffness. While traditional approaches to haptic modeling have relied on discrete classification or handcrafted physical parameters, this thesis advances a data-driven framework for predicting and synthesizing structured perceptual attributes across tactile and kinesthetic modalities. The goal is not merely to recognize object identity or detect contact events, but to model how users interpret continuous haptic qualities such as roughness, slipperiness, and mechanical realism.

For tactile perception, the thesis presents a framework to predict fine-grained perceptual ratings from multimodal interaction data. A structured four-dimensional perceptual space, based on psychophysical user studies, captures key texture qualities using bipolar dimensions (e.g., rough–smooth, flat–bumpy, sticky–slippery, hard–soft). To model this space, a dual-stream neural network is developed that combines visual texture features with tactile signals such as acceleration, force, and scanning speed. In addition to perceptual prediction, the thesis introduces the Fourier-enhanced Transformer Encoder Network (FoTEN) for real-time texture rendering. This model synthesizes high-frequency acceleration signals from interaction inputs by leveraging dedicated temporal and spectral encoders to preserve perceptual fidelity. Together, these components establish a bidirectional framework for perceptually aligned texture analysis and signal synthesis.

In the kinesthetic domain, the thesis tackles the challenge of aligning mechanical feedback from car door interactions with subjective user expectations. A structured perceptual vocabulary is derived from user experiments involving real vehicles and a programmable car door simulator. Using this dataset, a residual CNN-based model is trained to predict continuous ratings along seven bipolar perceptual dimensions, and an inverse decoder is developed to generate force profiles from

user-defined perceptual inputs. The system enables forward inference and reverse synthesis of door mechanics, supporting human-centered tuning and simulation without the need for physical prototypes. Perceptual experiments validate the consistency, controllability, and realism of the generated signals, particularly for physical attributes such as resistance and damping.

Across both tactile and kinesthetic settings, this thesis demonstrates how continuous, multi-dimensional haptic perception can be modeled and rendered using data-driven techniques. By moving beyond classification toward structured perceptual prediction and real-time synthesis, the work provides new tools for building haptic systems that adapt to and align with how users actually feel. These contributions open pathways for perceptual authoring in virtual environments, robotic manipulation, and haptic interface design.

# Acknowledgement

This Ph.D. journey has been both meaningful and challenging, and I would not have been able to reach this point without the grace of the Almighty. His guidance and blessings gave me the strength to keep going, especially during the most difficult times, and I am equally grateful to all those whose support and motivation helped me throughout the way.

I would like to sincerely thank my advisor, Prof. Dr. Seokhee Jeon, for his guidance and support throughout my doctoral studies. He has been patient and encouraging, offering valuable feedback and direction while allowing me space to grow at my own pace. His trust and advice helped me stay focused and develop both technically and personally over the course of this work. I am truly grateful for the opportunities he provided and the confidence he placed in me throughout this journey.

I am also thankful to all my colleagues in the Haptics and VR Lab at Kyung Hee University. The lab has been a place of learning, collaboration, and shared effort. I especially want to thank Dr. Waseem Hassan, Dr. Ahsan Raza, Ms. Tatyana Ogay, and other members, past and present, for their help and support throughout these years. Whether it was discussing ideas, solving problems, or simply sharing the daily routine, their presence made the journey smoother.

I am also grateful for the environment at Kyung Hee University, which provided a supportive space to work and grow. It allowed me to connect with researchers from various fields and learn from diverse perspectives. I would like to thank my friends and fellow Ph.D. students, including Mr. Abdul Muqeet, Mr. Salman Ali, and Dr. Junaid ur Rehman, for their support, thoughtful conversations, and shared experiences. Many others contributed in their own way, and although I cannot name everyone, I truly appreciate their presence and encouragement.

Most importantly, I would like to thank my family for being there for me throughout this journey. No words feel big enough for this part. My wife, Myrah Naeem, stood by me every single day. She managed so many things quietly behind the scenes while I was busy with work. Her patience, support, and belief in me made it possible to keep going. My son, Master Huraira, came into my life during this Ph.D. and brought happiness and balance, especially during stressful days. His presence always lifted my mood and helped me regain energy.

I am also deeply grateful to my parents, Muhammad Ibrahim Awan and Naghma Ibrahim, for their unwavering support, love, and prayers not only during my Ph.D. but throughout my entire life. Their belief in me has always been a source of strength, especially during difficult times. I am equally thankful to my siblings (Kanwal, Aisha, Mubeen and Palwasha) for standing by me and for taking care of our parents and many of my responsibilities back home while I was away. Knowing that my family was there gave me peace of mind and allowed me to stay focused on my work.

Although this thesis carries my name, it reflects the contributions, support, and kindness of many people. I am sincerely grateful to everyone who stood by me throughout this journey.

Mudassir Ibrahim Awan

August, 2025

# Table of Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

This thesis investigates how humans perceive tactile and kinesthetic interactions through physical contact and how such perceptual experiences can be modeled, predicted, and synthesized using data-driven approaches. The focus is on two representative scenarios: textured surface exploration for tactile perception and car door operation for kinesthetic perception. By bridging physical haptic signals with human cognitive impressions, this work aims to improve the realism and usability of haptic systems in applications ranging from virtual reality to product design.

## 1.1 Motivation

Touch enables humans to interpret and respond to the physical world through a combination of tactile and kinesthetic cues. Tactile perception involves high-frequency vibrations, frictional interactions, and fine surface features, while kinesthetic perception captures force, torque, and resistance associated with large-scale motion. These complementary modalities shape how people describe materials, assess quality, and interact with objects. For example, brushing a surface reveals roughness and slipperiness through vibrations, while opening a car door provides kinesthetic feedback that conveys weight or mechanical refinement. As haptic interfaces become more prevalent in virtual reality, teleoperation, and intelligent products, the need for perceptually aligned haptic modeling becomes increasingly important.

Despite technological advances in sensing and actuation, most current haptic systems remain disconnected from how humans actually interpret touch. Texture models are often selected without knowledge of how they feel, and kinesthetic responses are designed through physical tuning rather than perceptual objectives. A texture library may contain dozens or hundreds of modeled vibration signals, yet lack the perceptual labels that allow users to retrieve or synthesize content based on

how it feels. Similarly, designers of mechanical feedback systems, such as automotive doors, may wish to evoke sensations like “easy to pull” and/or “luxurious,” but are left adjusting torque profiles manually due to the absence of perceptual mappings.

What is missing is a unified framework that connects physical haptic signals with human perceptual interpretation. Such a framework should allow machines to predict how a signal will be perceived and to generate new signals based on perceptual targets. This would support both recognition and rendering of haptic content in a manner consistent with user expectations.

This thesis addresses this challenge through two application domains. First, for tactile perception of textured surfaces, it introduces (1) an attribute estimation model that predicts user-rated perceptual descriptors from multimodal sensor data and (2) a generative model for synthesizing texture-induced acceleration signals from interaction variables. Second, for kinesthetic perception of mechanical systems, it models the relationship between torque signals and perceived car door quality, supporting both prediction and generation of force profiles from perceptual attributes. Together, these contributions aim to support data-driven and perceptually grounded haptic systems across both modalities.

## 1.2 Research Problem

Most haptic modeling and rendering systems rely on raw physical signals such as acceleration or torque, with limited consideration of how those signals are interpreted by users. As a result, signal-based rendering often lacks perceptual meaning, and designers are forced to guess or manually tune haptic content.

Two key problems arise in this context:

- **In texture-based haptics:** A content creator may have access to a library of modeled texture signals, but cannot retrieve textures based on perceptual terms like “rough” or “soft” without attribute labels. Manually tagging each entry is impractical, and psychophysical labeling is often absent. There is also a need for scalable texture modeling techniques that generate realistic vibrations from input parameters like scanning speed and force.

- **In kinesthetic haptics:** An automotive designer may wish to render a car door that feels "luxurious" or "recoiling," but must rely on low-level signal tuning or physics simulation with no perceptual control. There is no mechanism to directly synthesize torque signals from user-defined adjectives, nor to predict how a given force profile will be experienced.

While prior work has explored signal-to-perception prediction, the reverse problem of generating haptic signals from perceptual input remains underdeveloped, particularly in kinesthetic systems. This limits interactive content design, simulation realism, and user-specific adaptation.

The core research problem is:

*"How can physical haptic signals be mapped to and from human perceptual attributes in a data-driven and scalable way, across both tactile and kinesthetic domains?"*

## 1.3 Contributions

To highlight the significance and scope of this research, the primary contributions of the thesis are outlined as follows:

**Perceptual Attribute Estimation for Textured Surfaces**

- Construction of Haptic Attribute Space:
  - Developed a perceptual attribute space through psychophysical user studies using a range of real-world textures.
  - Defined key bipolar descriptors such as rough to smooth, sticky to slippery, flat to bumpy, and hard to soft.
- Multimodal Dataset Collection:
  - Collected synchronized visual and tactile data for real textured surfaces, including acceleration signals, force, and scanning speed.
  - Acquired user ratings for each texture along the defined perceptual dimensions to serve as ground truth.
- Attribute Prediction Framework:
  - Proposed a deep learning framework combining visual and tactile features to estimate user-perceived attributes of textures.
  - Validated the model using leave-one-out cross-validation and demonstrated accurate estimation across a wide set of unseen textures.

**Modeling and Rendering of Virtual Haptic Textures**

- Data-Driven Texture Signal Modeling:
  - Proposed a deep learning–based framework to model the generation of tactile acceleration signals from interaction parameters such as speed and force.

  - Designed the architecture to capture temporal and dynamic properties of texture interactions for realistic reproduction.

- Tactile Signal Synthesis for Rendering:

  - Enabled generation of high-fidelity tactile signals that replicate real surface interactions in virtual scenarios.
  - Focused on the practical synthesis of vibration signals suitable for rendering through standard haptic actuators.

**Modeling and Generation of Car Door Perception**

- Force Profile Collection and Perceptual Space Construction:

  - Collected temporal force profiles from six distinct car door models under controlled experimental settings.
  - Conducted structured user studies using real doors and a programmable simulator to construct a perceptual space based on bipolar adjective pairs (e.g., light–heavy, smooth–jerky).
  - Refined the perceptual descriptors through iterative evaluation to ensure consistency and relevance across participants.

- Prediction of Perceived Qualities from Force Signals:

  - Developed a residual CNN-based encoder–decoder model to estimate user perception directly from time-series force signals.
  - Evaluated model accuracy using MAE, RMSE, and $R^2$ metrics across multiple perceptual dimensions.
  - Validated perceptual consistency through user studies on a high-fidelity car door simulator.

- Generation of Force Signals from Perceptual Attributes:

- Trained an inverse model that generates torque profiles based on specified perceptual inputs using a decoder-driven architecture.
- Enabled simulation of car door feel based on user-defined adjectives, supporting perception-driven tuning and virtual prototyping.
- Verified the generated signals through interactive user experiments, confirming alignment with intended perceptual impressions.

## 1.4 Thesis Organization

This thesis is organized into six chapters. Chapter 1 introduces the motivation behind the research, outlines the problem, and summarizes the main contributions. Chapter 2 reviews relevant literature in haptic perception, data-driven signal modeling, and haptic interfaces, providing the foundational context for the work. Chapter 3 presents the approach for estimating perceptual attributes of textured surfaces, including the construction of a perceptual attribute space, dataset collection, and a multi-modal prediction framework. Chapter 4 describes the modeling and rendering of tactile signals using a data-driven method to synthesize acceleration profiles from interaction parameters, along with quantitative and user-based evaluations.

Chapter 5 focuses on car door interaction and is broadly organized into three parts. The first part describes the collection of force profile data from real car doors, including data augmentation techniques and the construction of a perceptual space based on user experiments. The second part presents a method for predicting user-perceived attributes directly from force signals. The final part explores the inverse problem: generating force profiles that reflect user-defined perceptual attributes, offering a pathway toward experience-driven design of mechanical interactions. Chapter 6 concludes the thesis with a summary of findings and a discussion of future directions, including personalization and broader application of perception-based haptic modeling.

# Chapter 2

# Related Works

This chapter presents the theoretical and methodological foundations that support the objectives of this thesis. It provides an overview of perceptual modeling in haptics, data-driven techniques for signal estimation and generation, and the development of haptic systems for tactile and kinesthetic interaction. Each section introduces key concepts and research directions that form the basis for the methods and contributions proposed in later chapters.

## 2.1 Haptic Textures

### 2.1.1 Haptic Attributes and Perceptual Space

Haptic texture attributes characterize how surfaces are perceived through touch, encompassing qualities such as roughness, slipperiness, stiffness, and thermal feel. These attributes can be sensed through direct finger contact or tool-mediated interaction, forming the basis for constructing perceptual models of surface textures. Such models often rely on multidimensional spaces that capture tactile experiences from a user-centered viewpoint [1].

Early studies exploring tactile perception through unmediated contact include the work by Yoshida et al. [2], which identified four core perceptual scales: hardness, weight, temperature, and surface coarseness. Subsequent research refined these axes, with hardness and roughness emerging as the most dominant [3]. Further investigations contributed to a more nuanced understanding, introducing distinctions between macro- and micro-roughness, and highlighting friction and compliance as essential perceptual factors [4]. These efforts contributed to the identification of five primary perceptual components: macro-roughness, micro-roughness, friction, stiffness, and thermal properties.

In contrast, tool-mediated studies have examined how the use of rigid probes or styluses alters texture perception. LaMotte [5] showed that tapping or pressing with a stylus can enhance perception of certain attributes, particularly hardness and softness. Other research indicated that such tools are well-suited for capturing surface-level features like roughness while minimizing variability caused by skin deformation [1, 6]. However, finer aspects such as friction or micro-texture detail are often better perceived through direct contact [7].

To represent subjective impressions of texture, researchers have commonly employed dimensionality reduction methods such as Multi-Dimensional Scaling (MDS) [1] and Principal Component Analysis (PCA) [8,9]. These approaches reduce complex perceptual data to more manageable forms and are widely used in applications including surface classification [10, 11] and virtual texture design [12].

Although these techniques provide insights into the underlying structure of perception, they also introduce limitations. By compressing the dimensional space, they risk losing subtle but meaningful aspects of user ratings. In response to this, a recent approach proposed the use of a four-dimensional Haptic Attribute Space that preserves raw user input without applying projection-based simplification [13, 14]. This model retains full rating resolution and organizes the space into two intuitive two-dimensional subspaces, offering a more transparent and precise view of perceptual texture representation.

Despite significant advancements, further refinement of perceptual models is needed to fully reflect the complexity of user experience. Continued development in this area will support the design of tactile systems that more closely match the richness of human haptic perception.

### 2.1.2 Tactile and Vision Data for Texture Recognition

Texture analysis through tactile feedback involves capturing the characteristic vibrations generated during surface interaction. These signals encode both macro-level features, such as bumpiness, and micro-level details, including fine roughness [5, 7]. However, the mapping between texture properties, user interaction behavior, and the resulting vibrations is inherently nonlinear and complex, making accurate modeling and differentiation a challenging task [15]. Early studies addressed this by recording tactile data under fixed interaction conditions or by segmenting signals into station-

ary intervals, which limited the applicability and generalization of their models [16]. More recent work has adopted free-hand data collection strategies that preserve the natural dynamics of interaction and maintain the diversity of vibratory responses [17], minimizing information loss. Despite these advancements, both parametric and deep learning approaches continue to face difficulty in separating textures with similar vibratory profiles [17–19].

In parallel, vision-based approaches offer a less complex alternative to tactile sensing, with reduced hardware requirements. Heravi et al. [20] utilized GelSight imagery to perform effective texture classification, although their objective focused on rendering rather than accurate perceptual attribute estimation. Similarly, Yang et al. [21] aligned GelSight features with visual and auditory modalities to enhance classification. In contrast, Hassan et al. [14] used a handcrafted feature-based method with texture images to estimate haptic attributes, demonstrating strong performance overall but encountering difficulties with properties like softness and fine roughness. These limitations are likely due to the nature of visual features, which primarily capture macro surface patterns but fail to represent sub-surface characteristics such as compliance or softness, reducing their predictive accuracy in such dimensions.

The limitations of single-modality techniques have been well acknowledged in the haptics community. To address them, several studies have proposed combining visual and tactile modalities to achieve more reliable texture representations [8]. Visual inputs are effective in capturing macro structure, while tactile signals convey micro-level details, and their fusion creates a more complete encoding of surface properties. This multimodal integration has been shown to improve perceptual attribute prediction performance, going beyond basic classification tasks [22, 23]. By leveraging complementary patterns across both modalities, models benefit from enhanced generalization and increased robustness, even when encountering previously unseen textures. However, the majority of existing work remains focused on classification, with relatively few studies addressing direct regression of perceptual ratings for haptic attributes [22, 24].

### 2.1.3 Texture Perception Modeling with Deep Learning

Recent efforts have turned to deep learning for texture perception, blending visual and tactile data [19,23]. CNNs expertly extract spatial features from visual textures, nailing structural patterns

in images with top-notch texture recognition [14]. Meanwhile, tactile signals are inherently spatio-temporal, combining surface details and time-based changes.

Previous research employed convolutional neural network (CNN)-based models on time-series tactile signals to extract localized features from vibrational data [22]. With advancements in Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks which are crafted to capture temporal dependencies, there has been an investigation into hybrid CNN-LSTM frameworks. These frameworks typically integrate CNNs for the extraction of spatial patterns from input segments with LSTMs to learn the temporal progression across these segments [19]. Although this architecture facilitates the integrated modeling of spatial and temporal data, it also introduces challenges, including complex optimization processes, high sensitivity to hyperparameter tuning, and reduced spatial coherence resulting from sequential feature aggregation. Such limitations are not exclusive to haptic signal processing but have also been observed in various spatio-temporal learning tasks [19, 25, 26].

Recent research tackles these challenges by exploring new architectures for sequential data processing. Key approaches are Transformer models [27] and Convolutional LSTM (ConvLSTM) networks [28]. Though Transformers excel in sequence tasks, they often need a lot of data, which is less ideal for texture-based haptic uses with few samples [29]. On the other hand, ConvLSTM provides a structured, data-efficient way to capture both spatial and temporal relationships. It has shown great promise in various time-series tasks, like irrigation scheduling [30] and modeling surface deformation [31]. Its ability to maintain spatial structure while learning temporal patterns makes it ideal for handling dense sensor data such as tactile signals.

### 2.1.4 Acceleration-Based Texture Synthesis

The texture of a surface is shaped by its detailed micro and macro geometry. When exploring surfaces, either directly by hand or using tools, haptic texture perception emerges from intricate contact dynamics between the surface and the skin or tool. This interaction leads to variations in physical signals like displacement, force, and acceleration, which skin mechanoreceptors detect and the nervous system interprets as tactile texture.

Two primary methods exist for digitizing haptic textures. The first directly models the surface

geometry [32]. Despite being conceptually easy, this approach struggles with real-time rendering due to the computational load of simulating contact dynamics at over 1 kHz. Consequently, many use simplified assumptions to enable real-time rendering [33, 34].

The second approach avoids explicit simulation by recording physical signals generated during real surface interactions, along with associated interaction parameters. These signals are later interpolated based on user input to replicate the interaction during rendering. This method is particularly effective in tool-mediated settings, where surface contact generates high-frequency acceleration signals transmitted through a rigid tool. Texture modeling in this context becomes a matter of parameterizing, storing, and regenerating these acceleration profiles [11].

Recent studies focus on data-driven haptic texture modeling. Kuchenbecker et al. [35,36] captured vibrations from tool-surface contact. Acceleration profiles, mapped by force and speed, were interpolated to create new interaction signals, which were then transformed into high-frequency vibrations for texture feedback.

Abdulali et al. [37] extended this method by incorporating interaction direction into the parameter space, enabling more accurate rendering of anisotropic textures. Other studies have explored machine learning techniques. For example, Ujitoko et al. [38] trained a Generative Adversarial Network (GAN) to generate high-frequency acceleration signals from texture images, offering an image-based alternative for haptic feedback synthesis.

### 2.1.5 Deep learning for haptic texture modeling

In recent years, the focus of texture modeling moved to employing deep learning networks. The authors in [39] used a neural network approach for modeling and rendering textures. Their approach presented a significant advancement in accuracy, demonstrating the potential of neural networks in synthesizing haptic textures with fine accuracy. They used a motorized texture scanner, to collect data with different combinations of input variables. This data was then used for training their models and for rendering.

Subsequent research by the authors in [40] introduced a deep spatio-temporal network (DSTN) designed to align acceleration signals with near-constant speeds and forces. The DSTN architecture incorporated an attention-aware Convolutional Neural Network (CNN) stream for the extrac-

tion of spatial features, alongside a Bi-directional Long Short-Term Memory (Bi-LSTM) encoder-decoder network stream to capture temporal dependencies within the acceleration signals. While their approach demonstrated state-of-the-art performance, it potentially resulted in prolonged training and inference durations due to the extensive model size and the inherent recurrent nature of the LSTM. More recently, Culbertson's research team augmented the AR-based method by leveraging a Generative Adversarial Network (GAN) to synthesize novel textures [17]. Additionally, in [20], the authors introduced a unified, deep learning-based action-conditional model for haptic texture rendering, utilizing data derived from the vision-based tactile sensor (GelSight).

### 2.1.6 Transformers for time-series data modeling

Transformer-based networks have set new benchmarks in natural language processing tasks [27] and computer vision applications [41], outperforming traditional methods such as RNNs, LSTMs, and GRUs [42, 43]. While these traditional recurrent models are effective for process sequential data, they encounter significant challenges, including the vanishing gradient problem, limiting their ability to learn long-term dependencies in time series sequences and slow training speeds. Although LSTMs and GRUs incorporate mechanisms for selective memory retention and forgetting, their sequential processing nature still limits their effectiveness with long sequences [44]. The noteworthy success of transformers is predominantly attributed to their distinctive architecture, which facilitates the simultaneous processing of entire sequences, irrespective of the distances between different elements. This ability renders transformers especially adept at recognizing and assimilating recurring patterns with varying dependency lengths, thereby expediting training durations. Furthermore, recent investigations have explored and refined transformer-based methodologies for applications in time-series analysis, encompassing vibration signal classification [42, 45], music generation [46], and time series forecasting [47]. It is crucial to acknowledge that the application of transformer networks, originally comprising both encoder and decoder components, necessitates certain modifications contingent on the specific problem. Initially devised for sequence-to-sequence generation tasks such as translation [27], transformers utilize the encoder for processing input sequences and the decoder for generating corresponding output sequences. However, in scenarios where sequence generation is not the primary objective, as in classification, forecasting,

and regression tasks, employing only the encoder as a feature extraction module while excluding the decoder serves to simplify the model, enhance training efficiency, and mitigate overfitting. This streamlined method not only optimizes training time but also bolsters the model's robustness and generalization abilities, rendering it ideally suited for tasks requiring detailed feature analysis rather than sequence generation [44, 48].

## 2.2 Car Door

### 2.2.1 Machine Learning and Perception of Cars/Car Parts

Machine learning methodologies are employed in diverse elements of automotive design, including comfort, aesthetics, and usability [49]. By training models on extensive datasets that encompass information about car designs and user feedback, these investigations have succeeded in identifying patterns and relationships that can inform the design process. Notably, machine learning has demonstrated significant promise in predicting users' emotional responses to car designs. Researchers have developed models capable of accurately predicting users' emotional reactions to various car designs, based on characteristics such as color and shape [50]. This advancement has yielded insights into the emotional dimensions of automotive design and holds potential for informing the creation of vehicles that are more emotionally engaging. Furthermore, machine learning has been utilized to examine the relationship between the physical properties of cars and their perceived quality, such as the perceived quality of engine sound [51, 52]. Nevertheless, the use of machine learning to predict haptic perception and emotions associated with car doors remains a relatively unexplored area.

### 2.2.2 The Role of Emotions in Product Design

Emotions play a crucial role in shaping users' perception of products and their overall satisfaction [53]. Affective engineering has emerged as an interdisciplinary field that aims to incorporate users' emotions and preferences into the design process, thereby enhancing the overall user experience [54]. Some studies have explored the role of emotions in the context of automotive design, focusing on various aspects such as the interior environment, the driving experience, and the vehi-

cle's appearance [55]. However, there is still limited research on the role of emotions in the design of car doors and their impact on users' haptic perception and satisfaction.

Car door design plays a critical role in the overall user experience of a vehicle. Early research in this area focused on the optimization of car door dynamics, with an emphasis on improving the opening and closing characteristics [56]. This body of work has led to the development of various techniques and approaches for optimizing car door design, such as the use of advanced materials and manufacturing processes.

Recently, there has been a shift in focus toward understanding the relationship between car door design and perception. Studies have explored the impact of car door design on users' perception of quality and luxury [57]. These investigations have revealed that users associate certain design elements, such as the smoothness of the door opening and closing motion, with higher-quality vehicles. By adopting such an approach, designers can create car doors that not only perform well in terms of functionality but also evoke positive emotions and contribute to an overall satisfying user experience.

### 2.2.3 Haptic Perception in Automotive Design

Haptic perception, or the sense of touch, is integral to the way users experience and interact with products [58, 59]. Within the realm of automotive design, haptic perception includes not only the tactile sensations experienced upon contact with surfaces and materials [60] but also the kinesthetic feedback received from operating mechanisms [61, 62]. A more comprehensive understanding of haptic perception can aid designers in developing experiences that are both more satisfying and user-friendly [63]. Despite its significance, the exploration of haptic perception in automotive design remains limited, with few investigations examining the variables that influence the perception of car door quality and the emotions they elicit. By employing machine learning techniques, designers can develop more intuitive and engaging interfaces that accommodate the varied preferences and requirements of users.

### 2.2.4 Data-driven Approaches in Automotive Design

Data-driven approaches have gained traction in various fields, including automotive design, where they enable designers to make informed decisions based on empirical data [64]. Researchers have used data-driven models for various aspects of automobiles, such as improving the braking control systems [65], or evaluating the health of electronic systems on board [66].

Data-driven methodologies, in conjunction with machine learning techniques, can enhance the development of predictive models that elucidate the intricate relationships among product attributes [67]. Notwithstanding the prospective advantages, there remains a necessity for further investigation into data-driven strategies within the automotive design sphere, particularly concerning the perception of car doors and the emotions they evoke.

# Chapter 3

# Haptic Texture Attribute Estimation Using Visuo-Tactile Data

This chapter discusses the estimation of human-perceived texture attributes using a combination of visual and tactile sensor data. It presents the construction of a perceptual attribute space through psychophysical experiments and introduces a multi-modal learning framework designed to predict perceptual ratings from both visual and physical interaction signals.

## 3.1 Motivation

Human perception of textures arises from interpreting rich tactile cues, including surface-induced vibrations, applied force, and exploratory movement. These physical signals generate subjective impressions commonly described using perceptual attributes such as roughness, softness, and stickiness [18, 68, 69]. Organizing such descriptors into a structured perceptual space is fundamental for delivering haptic feedback that aligns with user expectations in virtual reality, robotic manipulation, and remote operation tasks [70].

To enable accurate perceptual modeling, this work defines a four-dimensional haptic attribute space, constructed through controlled psychophysical experiments. Fifty real-world textures are rated by participants along four bipolar dimensions: rough–smooth, flat–bumpy, sticky–slippery, and hard–soft [71,72]. Alongside these subjective ratings, physical measurements are collected for each texture sample. These include high-resolution visual images and time-synchronized tactile signals comprising acceleration, scanning speed, and applied force [22,73]. Establishing a reliable mapping from physical measurements to perceptual ratings is a critical step for systems that aim to predict user perception from sensory data.

Although earlier work has attempted perceptual attribute prediction using either vision or tactile sensing, unimodal models often struggle with generalization across diverse material properties and interaction scenarios [18, 19]. Tactile inputs provide temporally detailed information but are highly sensitive to interaction dynamics and sensor noise. In contrast, visual inputs capture stable macrostructure but cannot resolve fine properties such as compliance or surface friction [16, 74]. Because these two modalities encode complementary characteristics, their integration offers a more robust and complete representation of surface properties [73].

This chapter presents a deep neural architecture that jointly leverages visual and tactile signals to estimate perceived haptic attributes. The model adopts a dual-stream design: the visual branch extracts features using a convolutional autoencoder based on pre-trained ResNet embeddings [75] and gray-level co-occurrence matrix (GLCM) descriptors, while the tactile branch processes MFCCs computed from vibration signals along with speed and force data using a convolutional LSTM network [28]. Feature representations from both modalities are aligned and used to predict user ratings. The resulting model shows improved accuracy and generalization to novel textures [14, 16].

In addition to enhancing virtual interaction quality, this framework offers a scalable alternative to traditional psychophysical methods, which are often labor-intensive and costly [71, 72]. The proposed method also supports downstream applications such as perceptual compression and signal rendering, where transmitting only perceptual descriptors enables bandwidth-efficient reconstruction of realistic haptic experiences.

## 3.2 Overview

This section outlines the overall framework developed to predict perceptual haptic attributes from multimodal sensory input. The process begins by preparing texture samples drawn from a broad range of real-world materials. These samples are used to define two principal data representations. The first, referred to as the Physical Feature Space (PFS), comprises visual and tactile features extracted through systematic exploration using instrumented sensing tools. The second, the Haptic Perceptual Space (HPS), is constructed from user ratings obtained through psychophysical evaluation, capturing perceptual impressions along four bipolar dimensions [13].

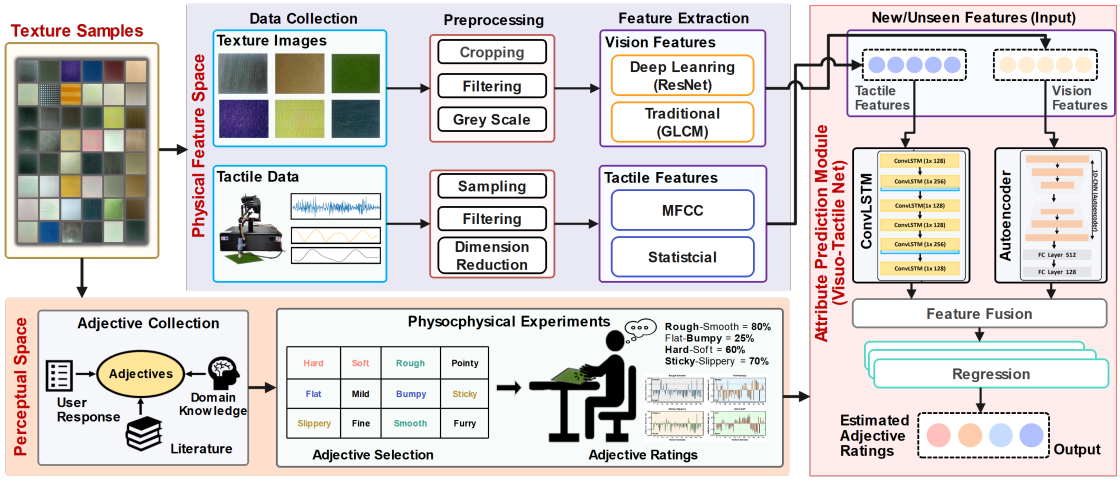


**Figure 3.1:** Schematic of the proposed attribute prediction framework. The model employs a dual-stream architecture to learn a mapping from visual and tactile features to perceptual ratings provided by users.

Central to the framework is the Attribute Prediction Module (APM), which models the relationship between the physical and perceptual domains. As shown in Figure 3.1, the APM employs a dual-stream neural architecture that integrates visual and tactile features to produce continuous predictions for each perceptual attribute. The specific design of this module is detailed in later sections. Data acquisition procedures and the composition of the PFS and HPS are elaborated in Sections 3.4 and 3.3, respectively.

## 3.3 Haptic Perceptual Space (HPS)

This section describes the construction of the Haptic Perceptual Space (HPS) based on psychophysical experiments with human participants. The process consists of two main stages. First, a preliminary study was conducted to determine a set of perceptual attributes that effectively characterize textural impressions. In the second stage, participants provided subjective ratings for each texture sample using the selected attributes. The experimental protocol and dataset were adapted from our prior work [14]. The following subsections summarize the key elements of this procedure.

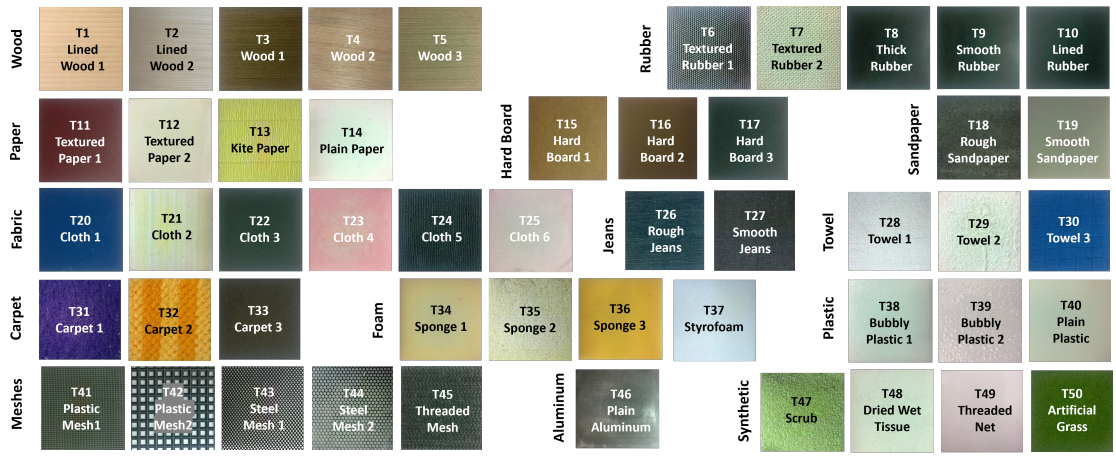


**Figure 3.2:** Fifty real texture samples used in this study, selected from a range of material categories to ensure diversity in tactile characteristics.

### 3.3.1 Texture Dataset

A total of 50 real-world texture samples were used to construct the Haptic Perceptual Space (HPS) and the corresponding Physical Signal Space (PSS). These textures were selected to represent a wide variety of surface types and material properties, enabling coverage across a diverse perceptual spectrum.

To ensure balanced representation, the textures were grouped into 16 categories based on material and tactile characteristics. These categories include wood, rubber, paper, hardboard, sandpaper, fabric, denim, towel, carpet, foam, plastic, mesh, aluminum, and other synthetic surfaces. The selected textures exhibit distinct levels of roughness, softness, friction, and compliance. A visual overview of the full dataset is provided in Figure 3.2.

Each sample was standardized to a size of 100 mm $\times$ 100 mm. The textures were mounted onto rigid acrylic plates of matching dimensions (100 mm $\times$ 100 mm $\times$ 5 mm) using liquid adhesive. This mounting approach ensured flat and consistent surface elevation across all textures, minimizing potential variations during user interaction [14, 19].

### 3.3.2 Experiment 1: Attribute Selection

The first stage of the psychophysical procedure aimed to identify perceptual adjectives that effectively characterize users' experiences when interacting with textured surfaces. A set of 60

**Table 3.1:** List of perceptual attributes shown to participants during the selection task. The four final bipolar pairs chosen for subsequent rating are emphasized in bold dark blue.

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

candidate attributes was compiled to represent the perceptual variations across the texture dataset.

These adjectives were collected from three primary sources: prior literature on haptic perception [68, 69], expert knowledge in the field, and preliminary user input. The complete list of selected adjectives presented to participants is provided in Table 3.1.

**Participants:** The study included 26 participants (19 males, 7 females) aged 25-34, with an average age of 28. All were right-handed, used their dominant hand for the tasks, and reported no disabilities affecting performance or needing special accommodations.

**Experiment Setup:** Figure 3.3 illustrates the experimental setup. Participants sat at a table and wore headphones emitting white noise to block outside sounds. In front of them was a cardboard box with two openings. One opening had a slit through which participants could insert their hands, preventing visual observation of the textures during exploration. The other opening enabled the experimenter to switch texture samples without exposing them. Instructions were given verbally and in writing to ensure participants understood the task, which was to choose descriptors that best represented the perceived texture qualities.

**Procedure:** This phase aimed to identify descriptive terms that effectively represent how textures feel to the touch. Participants interacted individually with 50 texture samples (see Figure 3.2) without time restrictions, freely using any tactile exploration method. For each texture, they chose adjectives from a precompiled list (Table 3.1) that best described their tactile perception. These

selections were recorded as binary values: ”1” if appropriate, ”0” otherwise. At the end of the session, participants suggested any additional adjectives that might better describe the textures.

**Results:** The aggregated adjective selections revealed consistent trends in how participants described texture samples. For each adjective, the binary selections were summed across all participants and samples, then normalized to yield a relevance score. Attributes with a relevance score of 50% or higher were retained, resulting in an intermediate set of 11 descriptors. From this subset, only adjectives with clearly defined antonyms were considered further. Antonymous pairs were constructed to represent bipolar perceptual scales. Attributes lacking direct opposites were removed. This process resulted in four final adjective pairs: rough–smooth, flat–bumpy, sticky–slippery, and hard–soft, which were used in the subsequent rating experiment.

### 3.3.3 Experiment 2: Adjective Ratings

**Experiment Setup:** In the second phase of the study, participants evaluated each texture sample using the antonymous attribute pairs selected in the previous experiment. A custom graphical interface displayed on a PC was used for recording responses (see Figure 3.3). The interface presented four sliders, each corresponding to one perceptual dimension.

The physical length of each slider was set to 127 mm, following established standards in perceptual scaling [76]. This configuration provided sufficient granularity for users to express nuanced perceptual differences while ensuring usability [76, 77]. Participants used their dominant hand to explore the textures at their own pace and adjusted the sliders accordingly. Each slider represented a bipolar attribute scale, with the verbal labels displayed at both ends. Numerical values ranged from 0 to 100 but were hidden from view during the task to minimize bias.

**Results:** Responses from all participants were aggregated to produce average ratings for each texture. For interpretability and visualization, the ratings were rescaled to range from -100 to 100, where 0 indicates a neutral point, and the endpoints correspond to opposing perceptual extremes (e.g., rough vs. smooth), with polarity denoting direction.

These averaged ratings form the final perceptual representation for each texture and serve as ground truth for learning the mapping from physical to perceptual domains, as discussed in

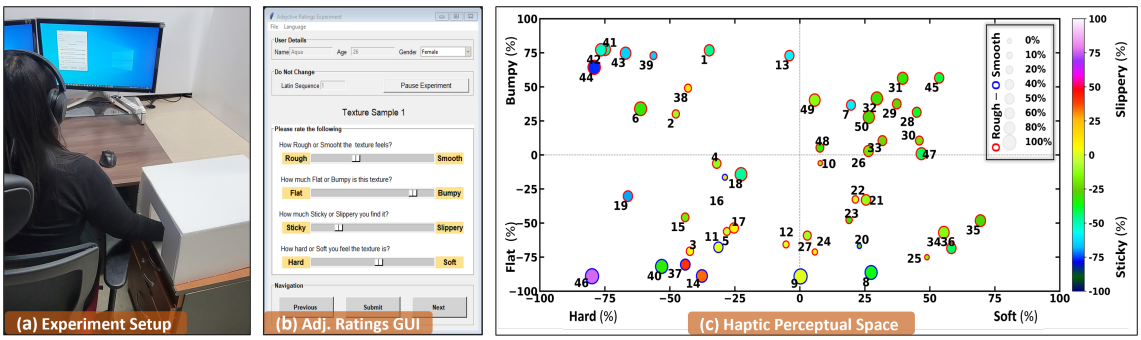


**Figure 3.3:** (a) Experimental setup used for perceptual adjective selection. (b) Graphical user interface (GUI) employed in the attribute rating task. (c) Visualization of haptic perception across four dimensions using a 2D bubble plot. Each point represents a texture, positioned by its perceived hardness (x-axis) and flatness (y-axis). The size of each bubble reflects roughness, while the color gradient encodes perceived stickiness. Ratings span from –100 to 100, capturing the full range between opposing adjective pairs (e.g., Flat to Bumpy).

Section 3.5. To visualize this multi-attribute dataset, we constructed a unified Haptic Perceptual Space (HPS) using a 2D bubble plot. In this visualization, the x-axis represents the hard–soft dimension, the y-axis corresponds to flat–bumpy, bubble size encodes rough–smooth, and the color gradient indicates sticky–slippery.

The resulting plot, shown in Figure 3.3, offers a compact and interpretable view of the four-dimensional perceptual ratings. To the best of our knowledge, this is the first visualization that consolidates multiple haptic dimensions into a single 2D framework with absolute rating values. Unlike previous approaches that depict each attribute in separate figures [14], this method provides a unified and efficient representation of texture properties, enabling intuitive analysis across the dataset.

## 3.4 Physical Feature Space

This section provides an extensive overview of the Physical Feature Space (PFS), which is a comprehensive dataset comprising synchronized tactile signals and visual inputs. The discussion begins with an in-depth examination of the tactile data, elaborating on the specifics of the hardware configuration employed for data acquisition, the methodologies implemented for signal capture, and the subsequent preprocessing steps that are undertaken. Additionally, this section delves into

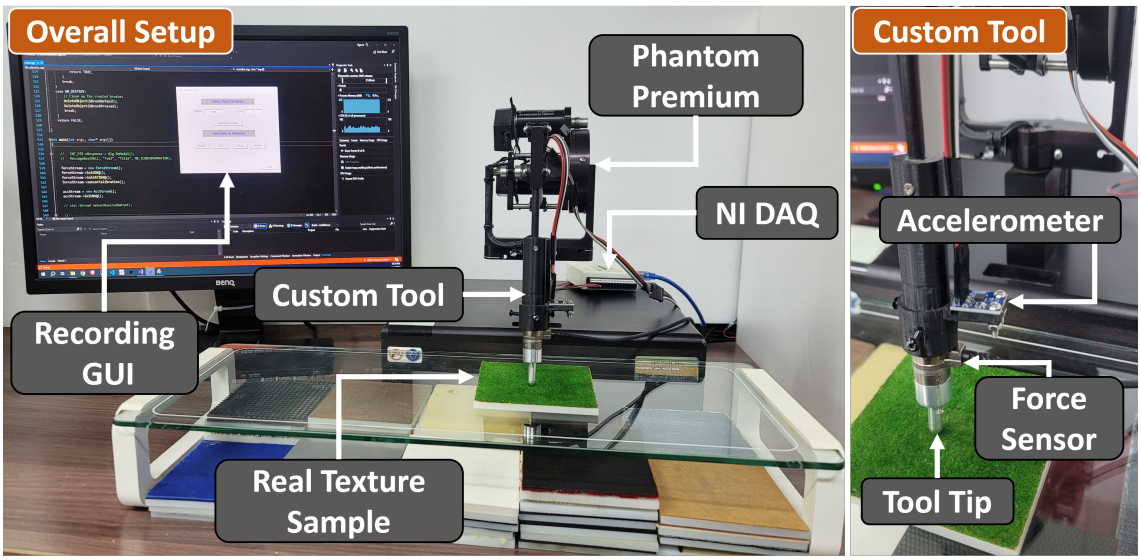


**Figure 3.4:** Tactile data acquisition setup. The system captures surface-induced vibrations, as well as the user's scanning speed and applied normal force during texture exploration.

the analysis of tactile features. Subsequently, attention is shifted towards the visual data component, where the process of image acquisition is thoroughly described. This includes an exploration of both deep learning and classical approaches to texture descriptor extraction, providing a well-rounded perspective on visual data processing.

### 3.4.1 Tactile Dataset

**Hardware Setup:** The arrangement for collecting tactile data, as illustrated in Figure 3.4, encompasses a robust, handheld instrument that incorporates a removable hemispherical steel tip with a diameter of 2.0 mm. The main structure of this tool is ingeniously crafted and produced through the use of 3D-printed ABS plastic, ensuring precision and durability. To capture any vibrations resulting from interactions with surfaces, an ADXL335, a 3-axis accelerometer developed by Analog Devices, is securely attached to the instrument. Simultaneously, a Nano17 force sensor from ATI Industrial Automation is employed to monitor three-dimensional force data. The integration of this tool with a Phantom Premium haptic device is critical, as it allows for precise tracking of both the tool's position and orientation. This integration significantly aids in accurately estimating the speed of interaction and the normal force applied, sampled meticulously at a frequency

of 1 kHz, enhancing the precision of the data collected. Furthermore, accelerometer readings are gathered via a National Instruments USB-6351 data acquisition module, operating at an advanced sampling frequency of 3 kHz. Meanwhile, a specialized data acquisition (DAQ) system is responsible for recording the force sensor outputs, achieving a notable sampling rate of 8 kHz. The configuration adheres strictly to the widely-recognized benchmarks in the domain of haptic texture analysis, ensuring high-resolution data that is indispensable for a thorough investigation of tactile signals [78].

**Data Collection and Pre-processing:**

Tactile data for all 50 textures (see Section 3.3.1) was collected through 60-second freehand exploration, allowing participants to interact naturally with each surface. To maintain consistency, all recorded signals were resampled to 1000 Hz. The first and last 2.5 seconds of each trial were trimmed to remove unstable segments at the start and end of the interaction.

Speed and force signals were smoothed using a low-pass filter at 25 Hz to eliminate high-frequency noise. Acceleration signals were filtered between 20 Hz and 500 Hz using a band-pass filter to retain relevant surface vibrations while removing drift and gravitational effects [11, 78]. To simplify the data, the three acceleration axes were combined into a single representative signal using the DFT321 algorithm, which preserves important vibrational patterns [79].

Scanning speed was calculated by combining velocity components from all three spatial directions, and normal force was derived by projecting the force vector onto the surface normal. Figure 3.5 shows an example of the final processed signals for the artificial grass texture (T50).

**Mel Frequency Cepstral Coefficients:** Acceleration signals recorded during texture interaction encapsulate substantial haptic information, accompanied by redundant components. To derive meaningful and concise features for haptic analysis, Mel Frequency Cepstral Coefficients (MFCCs) were utilized. Initially conceived for audio signal processing, MFCCs have demonstrated effectiveness in encapsulating the spectral characteristics of vibratory signals pertinent to surface classification and texture modeling [12, 15, 80].

To extract meaningful tactile features, the recorded acceleration signals were divided into 0.5-second segments, each containing 500 samples at a 1000 Hz sampling rate. A Hann window was

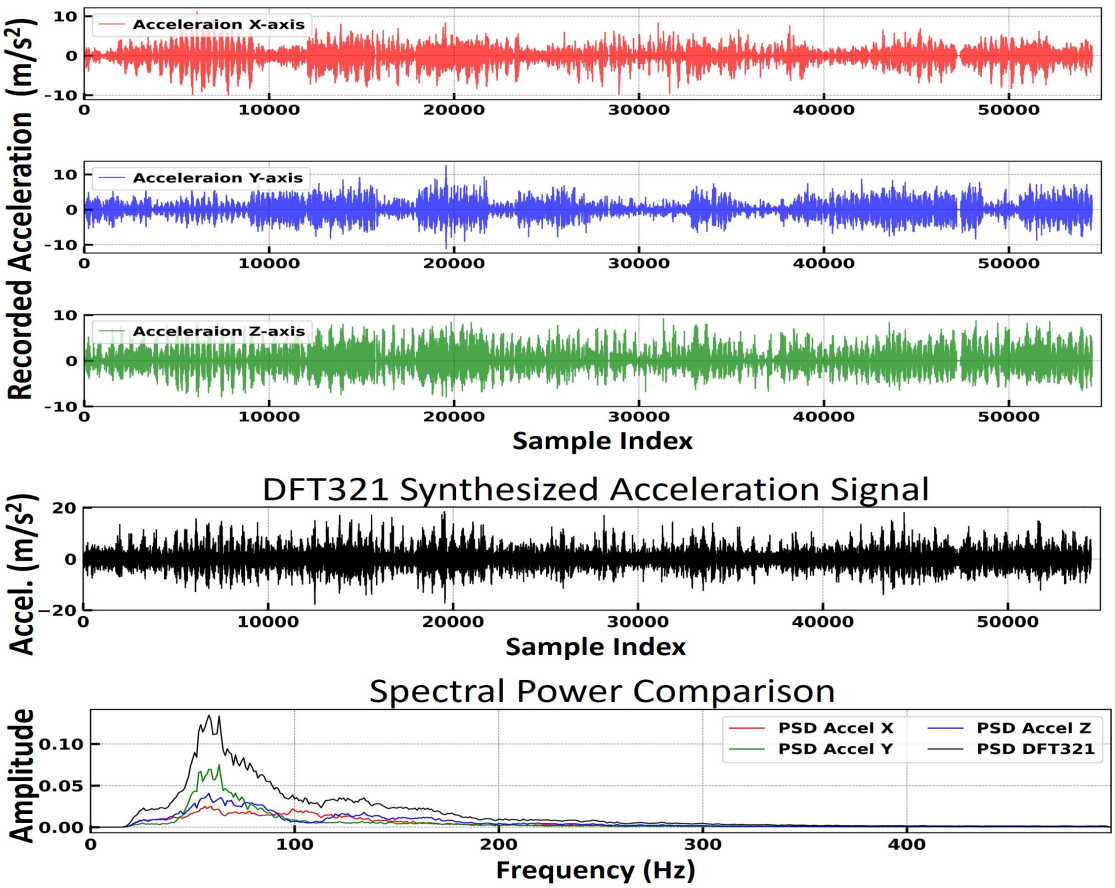


**Figure 3.5:** Acceleration signals captured for the artificial grass texture. The first three plots show raw signals along the x, y, and z axes. The fourth plot presents the unified signal obtained via the DFT321 algorithm, which preserves temporal structure. The fifth plot displays the corresponding spectral power distribution.

applied to each segment, which was then split into overlapping frames of 25 milliseconds with 50% overlap, yielding 40 frames per segment. From each frame, 13 Mel Frequency Cepstral Coefficients (MFCCs) were computed, resulting in a $40 \times 13$ matrix. This matrix was flattened into a single 520-dimensional vector to serve as the representation of vibration characteristics.

Unlike acceleration, the scanning speed and normal force exhibited relatively stable behavior over short durations. Therefore, for each 0.5-second segment, the minimum, maximum, and mean values of both signals were calculated, contributing six additional features. Together, the MFCC-derived and statistical features formed a comprehensive 526-dimensional feature vector for each segment. The feature extraction process was implemented using Python libraries, including SciPy

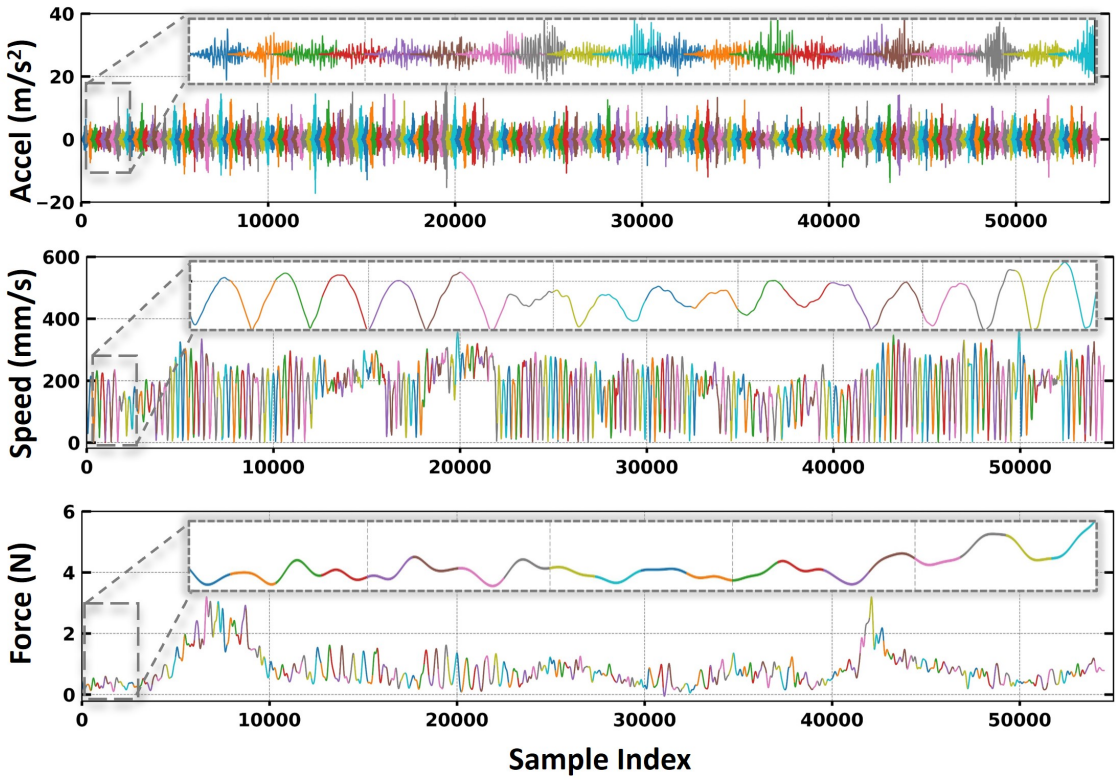


**Figure 3.6:** Processed acceleration, scanning speed, and normal force signals. Each plot includes segmentation boundaries generated during the preprocessing stage.

and librosa. These features were then fed into the tactile branch of the proposed model (HT-Net). A visual example of segmented signals for the artificial grass texture (T50) is presented in Figure 3.6.

### 3.4.2 Image Dataset

The multimodal methodology employed in this research integrates visual data and tactile signals to augment the prediction of haptic attributes. Historically, visual cues have been utilized in texture analysis through classical descriptors such as the Gray-Level Co-occurrence Matrix (GLCM), Gabor filters, and Local Binary Patterns (LBP) [22], as well as through more contemporary deep learning-based feature extractors [23]. Despite the fact that deep neural networks yield robust representations, they may be inadequate in capturing fine surface details when texture characteristics substantially deviate from the training data. To mitigate this issue, a hybrid feature extraction

strategy was implemented, incorporating both handcrafted and learned visual descriptors.

**Image Capturing Setup:** Capturing high-resolution visual data is essential for extracting meaningful features that relate to human haptic perception. To build a robust visual dataset of surface textures, a dp2 Quattro SIGMA digital camera was used, selected for its exceptional optical clarity and fine detail reproduction. The camera was mounted on a fixed tripod to maintain a consistent distance and angle across all samples. Each texture was photographed from a vertical height of 30 centimeters, ensuring standardized scale and perspective throughout the dataset.

To increase the diversity and realism of the captured images, each of the 50 texture materials was imaged 10 times under systematically varied lighting conditions. This included changes in illumination direction, intensity, and ambient light settings. The goal was to simulate different real-world viewing scenarios and introduce controlled variability in texture appearance, thereby enhancing the model's ability to generalize across unseen lighting conditions.

Following image capture, all photographs were centrally cropped to remove background elements and emphasize the main texture region. They were then resized to uniform dimensions of $1568 \times 1568$ pixels. This standardization step ensured consistent framing across all samples while eliminating potential artifacts near the borders. The resulting image dataset provided a rich visual representation of the textures, suitable for downstream feature extraction using both deep learning and statistical methods.

**DL-Based Features:**

To obtain meaningful deep visual descriptors from surface textures, ResNet-50 [75] was employed due to its robust architecture with residual connections and its proven effectiveness in material and texture classification applications [14,23]. Prior to feature extraction, each high-resolution texture image was divided into 49 partially overlapping patches, each measuring $224 \times 224$ pixels, to align with the input size required by ResNet-50. This patch-wise division allowed localized feature extraction across the surface.

For each patch, activations were collected from the average pooling layer, yielding a 2048-dimensional feature vector. These vectors were then aggregated by computing their mean, resulting in a single representative descriptor for the entire image. Since color information contributes

minimally to haptic texture perception, all images were first converted to grayscale. To maintain compatibility with the pre-trained ResNet architecture, each grayscale image was duplicated across three channels, preserving texture structure while fulfilling the model's RGB input expectations.

**Classical Texture Descriptors:** In parallel with the deep feature extraction process, classical texture descriptors were obtained using the Gray Level Co-occurrence Matrix (GLCM) method [81]. Each texture image was first quantized into 16 discrete gray levels, allowing for the construction of a $16\times16$ co-occurrence matrix that captured the spatial distribution of intensity transitions. This matrix was then reshaped into a 256-dimensional vector representing second-order statistical texture patterns. The resulting GLCM features were concatenated with the 2048-dimensional ResNet-derived features, forming a comprehensive 2304-dimensional visual representation for each texture sample.

To improve generalization and ensure consistency with the temporal variability present in tactile signals, data augmentation was applied during training using TensorFlow's built-in preprocessing tools. Augmentation procedures included random rotations, horizontal and vertical flipping, and the addition of Gaussian noise, thereby introducing controlled variability and enhancing the robustness of the visual encoding pipeline. This composite feature vector served as the input to the Haptic Vision Network (HV-Net), forming one stream of the visuo-tactile attribute prediction framework.

## 3.5 Attribute Prediction Module

To model the relationship between physical interaction data and subjective haptic perception, a dual-stream architecture referred to as the Visuo-Tactile Network is employed (see Figure 3.7). This framework comprises two specialized branches: the Haptic Vision Network (HV-Net), which encodes visual texture cues, and the Haptic Tactile Network (HT-Net), which captures temporal patterns from tactile signals. Instead of relying on raw input modalities, both networks operate on curated feature representations, a design choice that improves generalization performance and mitigates overfitting, as outlined in Sections 3.4 and 3.3.

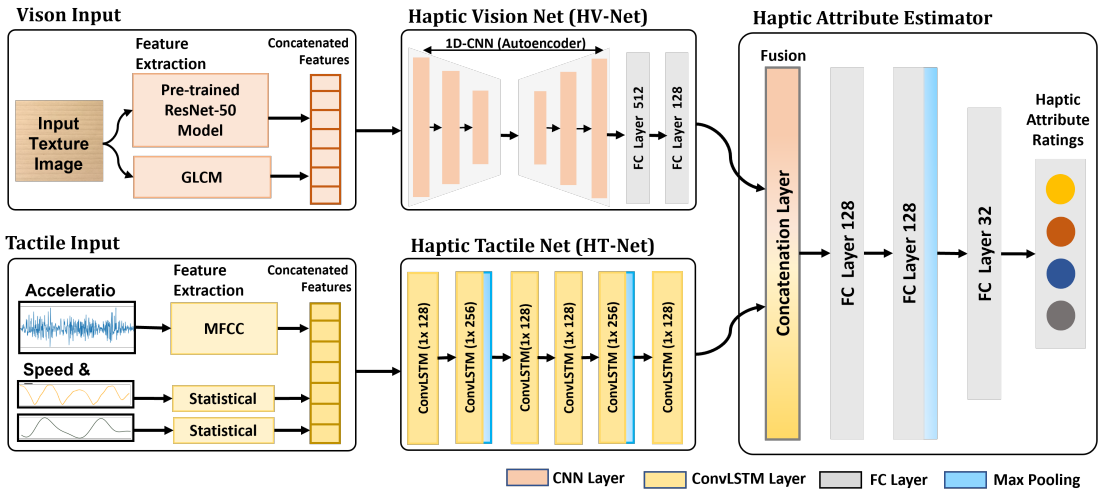


**Figure 3.7:** Architecture of the proposed visuo-tactile network. The model comprises two parallel streams: a visual branch implemented with a convolutional autoencoder and a tactile branch based on a 1D convolutional network.

The visual stream focuses on extracting spatial and statistical texture patterns from images, whereas the tactile stream processes acceleration, speed, and force signals to reflect dynamic surface interactions. Feature vectors from both branches are fused into a multimodal representation that encodes complementary information from vision and touch. This fused representation serves as the input for perceptual attribute regression. Detailed architectural components and training procedures for each stream are presented in the following subsections.

### 3.5.1 Haptic Vision Network (HV-Net)

The Haptic Vision Network (HV-Net) processes visual texture features through a one-dimensional convolutional autoencoder, tailored to handle high-dimensional non-spatial inputs while promoting generalization. The encoder is composed of five 1D convolutional layers, where the first two layers apply 256 filters with a kernel size of $(1 \times 4)$, followed by three layers with 128, 64, and 32 filters, each using a $(1 \times 3)$ kernel. Each convolutional layer is followed by a max pooling operation with a $(1 \times 2)$ window to reduce resolution and increase robustness. This design enables the encoder to progressively compress the input while capturing local feature patterns.

The decoder follows a symmetric architecture, reversing the transformation by gradually reconstructing the input using five 1D convolutional layers with filter sizes increasing from 32 to

256. This reconstruction task encourages the network to learn stable and informative representations by focusing on meaningful structural features in the visual data.

After reconstruction, the output is passed through two fully connected layers with 512 and 128 units, respectively, using ReLU activation. The final output is a 128-dimensional compact feature vector that captures essential visual texture characteristics and is later fused with the tactile stream for multimodal haptic attribute prediction.

### 3.5.2 Haptic Tactile Network (HT-Net)

Surface interaction generates tactile responses driven by user-modulated variables such as scanning speed $(v)$ and applied force $(f)$. These responses are encoded in acceleration signals $(a)$, which reflect surface microstructure and frictional characteristics. However, such signals can be influenced by external noise and variability in exploration patterns. To address this, HT-Net operates on features extracted from segmented time windows rather than using raw signals directly.

For each segment, MFCC features are extracted from the acceleration data, capturing the vibratory content in a compact and noise-resilient representation. Meanwhile, the speed and force signals, which contain lower-frequency components and exhibit greater temporal stability, are summarized using statistical metrics. The full segment-level feature vector is defined as:

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

where $\text{MFCC}_a$ denotes cepstral coefficients computed from acceleration, and the statistical terms represent aggregated metrics (minimum, maximum, average) from $v$ and $f$. Each $X_t$ vector corresponds to a single time segment.

To model sequential dynamics, a ConvLSTM-based architecture is employed. ConvLSTM layers integrate convolutional operations with recurrent memory, allowing the network to learn both local temporal patterns and long-range dependencies [28]. The architecture includes six stacked 1D-ConvLSTM layers with filter sizes of 128, 256, 128, 128, 256, and 128. Temporal downsampling is performed using max pooling layers with window size $1 \times 2$ placed after the 2nd, 4th, and 5th layers to reduce sequence length while retaining relevant features.

The final hidden state of the last ConvLSTM layer serves as the tactile representation, yielding

a 128-dimensional vector. This output is then combined with the visual embedding from HV-Net to form a unified feature representation for attribute regression. All layers were implemented using TensorFlow Keras [82].

### 3.5.3 Output and Training Method

The final visual and tactile feature vectors obtained from HV-Net and HT-Net, each comprising 128 dimensions, are concatenated to form a joint 256-dimensional multimodal representation. This vector is subsequently processed through two fully connected layers, each containing 128 neurons and activated using the ReLU function. To reduce dimensionality and enhance generalization, a max pooling operation with a pooling window of size $1 \times 2$ is applied to the output of the second dense layer. The pooled features are then passed through an additional fully connected layer with 32 neurons, followed by a final output layer consisting of 4 neurons that produces continuous predictions corresponding to the four perceptual attribute scores.

The complete architecture was empirically optimized through iterative validation experiments, including tuning the number of layers, filter widths, and neuron counts. Training is performed end-to-end using the TensorFlow-Keras framework, with the Adam optimization algorithm and root mean squared error (RMSE) as the objective loss function. ReLU activations are applied to all intermediate layers, while the output layer uses a linear activation to support real-valued regression outputs. The model is trained for a maximum of 200 epochs, with early stopping based on validation loss monitored across epochs. A patience threshold of 10 epochs is employed to prevent overfitting and ensure convergence stability.

## 3.6 Evaluation Experiments

Here we details how the proposed system was tested for its ability to infer human-perceived attributes from textured materials. It outlines the evaluation strategy, including the specific metrics used for error analysis and a cross-validation routine in which each texture is excluded once to test generalizability. Results are presented across multiple configurations, allowing for both quantitative benchmarking against alternative techniques and an examination of how different input

features influence prediction quality.

### 3.6.1 Error Metrics

### 3.6.2 Evaluation Metrics

To assess the predictive accuracy of the proposed framework, two commonly used evaluation metrics were adopted: Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). MAE reflects the average magnitude of absolute differences between the predicted and actual values, providing a direct measure of consistency. RMSE incorporates squared error terms, assigning greater weight to larger deviations, and thus captures both variance and precision of the predictions. These metrics were used to evaluate agreement between model outputs and user ratings for the four bipolar haptic attribute pairs: Rough–Smooth, Flat–Bumpy, Sticky–Slippery, and Hard–Soft, as defined in Section 3.3. The equations for each metric are as follows:

$$\text{MAE} = \frac{1}{N}\sum_{j=1}^{N}|\hat{r}_j - r_j|, \tag{3.1}$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{j=1}^{N}(\hat{r}_j - r_j)^2}, \tag{3.2}$$

where $\hat{r}_j$ denotes the model-predicted rating for the $j^{\text{th}}$ texture sample, $r_j$ is the corresponding user-provided rating, and $N$ is the total number of samples evaluated.

To ensure consistent interpretation, all ratings were normalized to a 0 to 100 range before computing the error values. In this normalized space, an MAE of 10 indicates an average prediction deviation of 10 units, directly reflecting the level of perceptual mismatch between model estimates and user judgments.

### 3.6.3 Evaluation Technique: Leave-One-Out Cross Validation

### 3.6.4 Cross-Validation Strategy

Reliable evaluation is crucial when working with datasets that are high in dimensionality yet limited in sample size. In such cases, effective validation strategies help prevent overfitting and ensure

that the model generalizes well to unseen data. One widely adopted technique is cross-validation, which involves dividing the dataset into subsets that are used iteratively for training and validation [83].

Among various schemes, $k$-fold cross-validation is commonly used, where the dataset is split into $k$ equal portions. The model is trained on $k-1$ folds and validated on the remaining fold, repeating this cycle $k$ times to obtain an average performance estimate. Although this method balances computational efficiency with reasonable coverage, using small values of $k$ may not fully capture the data variability, especially in smaller datasets.

To address this limitation, the leave-one-out cross-validation (LOOCV) approach sets $k$ equal to the number of samples $n$, evaluating the model using each data point individually as a validation case. For every iteration, the model is trained on $n-1$ samples and tested on the one left out. This cycle is repeated until each texture has been used once as the test instance [83–85].

In the present work, LOOCV was performed across all 50 texture instances, resulting in 50 training–testing cycles. Each fold offered an independent measure of prediction accuracy on a single unseen texture. Although computationally demanding, this method ensured comprehensive use of the dataset and provided robust and unbiased estimates of model generalization for texture attribute prediction.

### 3.6.5 Model Performance

Figure 3.8 illustrates the comparison between the predicted perceptual ratings generated by the proposed visuo-tactile network and the ground truth ratings provided by users. Each prediction is shown for all 50 textures, with values normalized within the range of -100 to 100.

The plot demonstrates a strong correspondence between predicted and actual ratings across the majority of textures, indicating that the model effectively captures the perceptual characteristics encoded in the visuo-tactile features. The alignment suggests that the dual-stream network generalizes well to unseen textures and maintains consistent prediction quality across different perceptual dimensions.

To further evaluate prediction accuracy, the Mean Absolute Error (MAE) was computed for each perceptual attribute. Among the four dimensions, the Flat–Bumpy (F–B) attribute yielded the

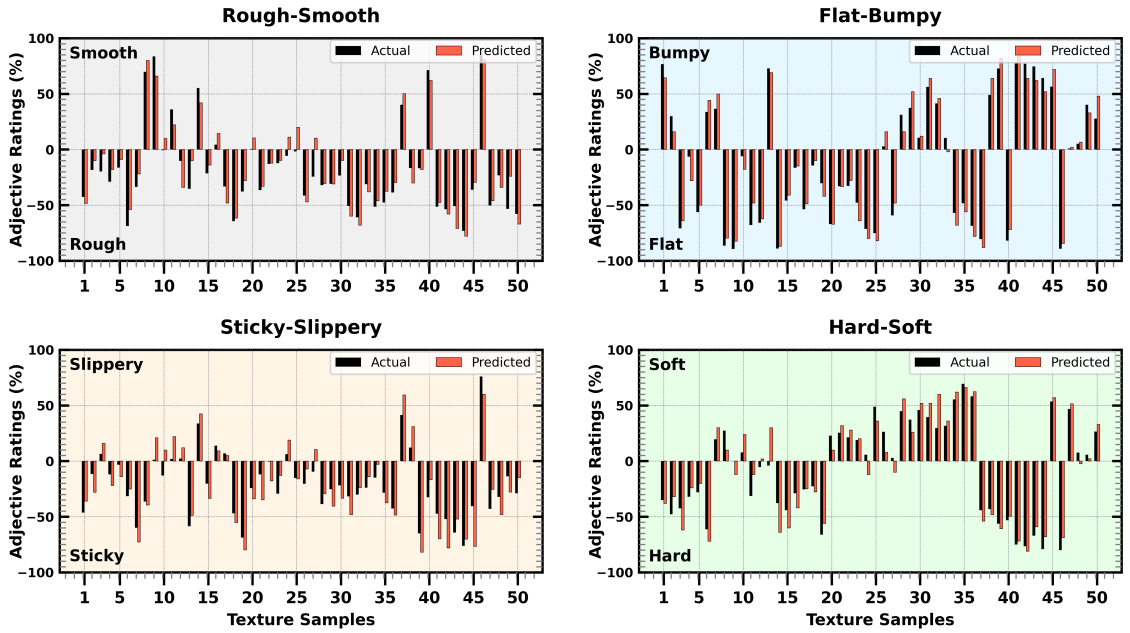


**Figure 3.8:** Predicted versus actual perceptual ratings for all 50 textures, evaluated using the Leave-One-Out Cross-Validation (LOOCV) method.

lowest error at 4.48, followed by Hard–Soft (H–S) and Rough–Smooth (R–S), with MAE values of 5.21 and 5.23, respectively. The highest error was observed for the Sticky–Slippery (S–S) dimension at 6.67, as summarized in Table 3.2. All reported MAE values are scaled to a 0–100 range, as outlined in Section 3.6.1, enabling direct comparison across attributes.

In addition, class-wise error distribution is presented in Figure 3.9. The results indicate that texture classes such as paper and denim exhibited relatively higher prediction errors, whereas other classes showed more consistent and accurate estimations. A more detailed analysis of these trends is discussed in Section 3.7.

### 3.6.6 Comparison with Baseline Models

In order to evaluate the efficacy of the proposed visuo-tactile framework, its performance was systematically compared with a selection of extant models specifically designed for the task of estimating haptic attributes through visual and/or tactile data. These models encompass the Vision 1D-CNN [86], Haptic CNN [14], Tactile CNN-LSTM [19], Tactile SVM [87], and a multimodal artificial neural network (ANN) baseline. Each model was implemented utilizing TensorFlow

**Table 3.2:** Comparison of Mean Absolute Error (MAE) across four perceptual attribute pairs for the proposed method and five baseline models.

| **Methods** | **R-S** | **F-B** | **S-S** | **H-S** |
|---|---|---|---|---|
| Artificial Neural Network | 21.13 | 26.12 | 22.85 | 25.44 |
| Vision 1D-CNN [86] | 18.55 | 19.63 | 17.89 | 17.24 |
| Haptic CNN [14] | 13.17 | 11.32 | 12.01 | 8.38 |
| Tactile CNN-LSTM [19] | 10.58 | 8.98 | 13.76 | 11.92 |
| Tactile SVM [87] | 9.40 | 14.89 | 15.35 | 10.54 |
| **Proposed Method** | **5.23** | **4.48** | **6.67** | **5.21** |

**Table 3.3:** Root Mean Square Error (RMSE) comparison across four perceptual attributes for the proposed framework and five baseline models.

| **Methods** | **R-S** | **F-B** | **S-S** | **H-S** |
|---|---|---|---|---|
| Artificial Neural Network | 24.41 | 31.62 | 25.73 | 32.19 |
| Vision 1D-CNN [86] | 22.35 | 24.88 | 19.61 | 20.59 |
| Haptic CNN [14] | 18.21 | 12.15 | 14.19 | 12.65 |
| Tactile CNN-LSTM [19] | 13.45 | 10.65 | 15.20 | 13.78 |
| Tactile SVM [87] | 11.26 | 16.37 | 20.81 | 11.93 |
| **Proposed Method** | **6.81** | **5.67** | **7.52** | **6.13** |

2.7, with their architectures faithfully reproduced based on the specifications detailed in their respective publications. To ensure fairness in the evaluative process, the output layer of each model was modified to regress four continuous variables corresponding to the predetermined perceptual dimensions.

The ANN baseline serves as a reference multimodal model without specialized structural modeling. It utilizes the same input features as the proposed system: visual descriptors combining ResNet and GLCM features, and tactile inputs composed of MFCCs and statistical measures. These features are fed into two modality-specific branches, each composed of four fully connected layers (128, 256, 256, 128 units). The outputs from both branches are then fused and processed by two additional fully connected layers with 64 neurons each, followed by a final output layer predicting the four attribute scores.

This comparison enables an evaluation of the contribution made by modality-specific architectures and structured temporal modeling in improving attribute prediction accuracy.

The performance results presented in Table 3.2 (MAE) and Table 3.3 (RMSE) confirm that

the proposed visuo-tactile framework consistently achieves superior accuracy across all four perceptual dimensions. The model yielded the lowest error values among all tested approaches, indicating strong generalization and precise attribute prediction. Specifically, the proposed method produced MAE values of 5.23 for Rough–Smooth (R–S), 4.48 for Flat–Bumpy (F–B), 6.67 for Sticky–Slippery (S–S), and 5.21 for Hard–Soft (H–S). In comparison, the ANN baseline resulted in notably higher MAE values, such as 21.13 for R–S and 25.44 for H–S. A similar pattern emerged in RMSE, where the proposed model achieved the lowest errors: 6.81 (R–S), 5.67 (F–B), 7.52 (S–S), and 6.13 (H–S). By contrast, the ANN baseline reported substantially larger RMSE values, including 24.41 (R–S) and 29.12 (H–S).

Among the baseline techniques, both the Tactile SVM [87] and CNN-LSTM [19] demonstrated improved performance over ANN but still fell short of the proposed approach. For example, in the F–B dimension, the proposed method achieved a MAE of 4.48, substantially outperforming the Vision 1D-CNN model [86], which recorded an error of 18.55. This trend was mirrored in RMSE, with Vision 1D-CNN yielding 24.85 versus 5.67 for the proposed model.

In summary, vision-only models [14, 86] consistently underperformed compared to tactile and multimodal methods. These findings highlight the critical role of tactile signals in capturing perceptual nuances and underscore the advantage of combining visual and tactile modalities for accurate haptic attribute estimation.

To further examine the contribution of model architecture independently from the input features, an additional analysis was conducted by retraining the baseline models using the same visual and tactile feature sets employed in this study. These include ResNet and GLCM features for visual

**Table 3.4:** Root Mean Square Error (RMSE) comparison of baseline models using the visual and tactile feature sets defined in this study.

| Feature Type | Methods | R-S | F-B | S-S | H-S |
|---|---|---|---|---|---|
| Vision | Vision 1D-CNN [86] | 28.76 | 34.32 | 25.54 | 31.76 |
| | Haptic CNN [14] | 21.5 | 13.82 | 12.89 | 15.48 |
| | **Haptic Vision Net (Ours)** | **13.26** | **10.11** | **12.52** | **8.6** |
| Tactile | Tactile CNN-LSTM [19] | 11.74 | 9.80 | 13.86 | 11.66 |
| | Tactile SVM [87] | 18.38 | 27.38 | 32.35 | 36.96 |
| | **Haptic Tactile Net (Ours)** | **9.89** | **11.35** | **10.71** | **7.98** |

input, and MFCC-based descriptors for tactile input. The RMSE values are presented in Table 3.4. The results show that while a few models such as Haptic CNN showed partial improvement with the proposed features, others like Vision 1D-CNN performed worse. In contrast, the proposed Haptic Vision Net and Haptic Tactile Net achieved lowest RMSE across all four attribute pairs. For example, Haptic Vision Net achieved an RMSE of 13.26 for R-S compared to 21.5 by Haptic CNN using the same features. Similarly, in the tactile domain, the proposed architecture outperformed CNN-LSTM and SVM models. These findings indicate that the performance improvement is not solely due to better feature design but also results from the structured, modality-specific architecture that captures spatial, temporal, and spectral information more effectively.

### 3.6.7 Individual Feature Error

This subsection explores the contribution of individual feature sets from both visual and tactile modalities to the enhancement of haptic attribute prediction performance. The objective of the analysis is to identify which feature types are most effective in reducing estimation errors, as well as to assess the additional benefits derived from integrating multiple modalities. For the visual modality, feature sets derived from ResNet-50 and the Gray-Level Co-occurrence Matrix (GLCM) were examined [14, 17]. On the tactile front, three signal representation methods were analyzed: 1D Discrete Wavelet Transform (1D-DWT), Discrete Fourier Transform (DFT), and Mel-Frequency Cepstral Coefficients (MFCC). These techniques were chosen based on their established efficacy in capturing pertinent patterns for haptic perception and texture classification [20, 22, 88].

Table 3.5 illustrates the performance of both individual and combined features within the domains of visual and tactile modalities. For inputs derived from vision, the concatenation of ResNet and GLCM features resulted in enhanced accuracy across all evaluated attributes. The amalgamated visual features attained an RMSE of 10.11 for R-S, surpassing the individual performance of ResNet (18.29) and GLCM (19.11). Parallel enhancements were noted for F-B and S-S. In the tactile domain, MFCC consistently demonstrated superior performance over 1D-DWT and DFT, achieving an RMSE of 9.89 for R-S as compared to 31.3 and 34.61, respectively. Significantly, due to the suboptimal performance of DWT and DFT, their integration with MFCC was not pursued,

**Table 3.5:** Root Mean Square Error (RMSE) comparison of individual feature types versus combined feature representations.

| 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** |

as preliminary trials yielded unstable and degraded results. The integration of visual and tactile features further minimized errors, yielding the lowest RMSE across most attributes. The proposed model, which integrates ResNet, GLCM, and MFCC, attained RMSE values of 6.81 for R-S and 5.67 for F-B. The tactile data alone also surpassed the performance of vision-only features for F-B, underscoring the significance of tactile input in specific perceptual dimensions. In summary, the results underscore the efficacy of multi-feature, multimodal fusion in enhancing the prediction of attributes.

Table 3.5 summarizes the performance of individual and combined feature sets across visual and tactile modalities. Within the visual domain, merging ResNet and GLCM features consistently improved accuracy across all attributes. For instance, the combined visual representation yielded an RMSE of 10.11 for the R–S attribute, outperforming ResNet alone (18.29) and GLCM alone (19.11). Similar improvements were observed for the F–B and S–S dimensions. In the tactile domain, MFCC features showed superior performance relative to both 1D-DWT and DFT, achieving an RMSE of 9.89 for R–S, compared to 31.30 and 34.61, respectively. Given the comparatively poor performance of DWT and DFT, they were not combined with MFCC, as preliminary tests indicated unstable results and performance degradation. Fusing visual and tactile features led to additional reductions in error, with the integrated model combining ResNet, GLCM, and MFCC achieving the lowest RMSE values, such as 6.81 for R–S and 5.67 for F–B. Notably, tactile-only features outperformed vision-only inputs for the F–B dimension, underscoring the critical role

of tactile information for certain perceptual attributes. These results validate the advantage of combining multiple feature types and modalities to enhance prediction accuracy.

## 3.7 Discussion

Drawing from the results presented in Figure 3.8 and Table 3.2, this section discusses trends in prediction accuracy across perceptual attributes, differences in modality-specific behavior, and error patterns observed at the texture class level. Among the four attribute pairs, the Sticky–Slippery (S–S) dimension resulted in the highest prediction error, while the Flat–Bumpy (F–B) attribute showed the lowest. The Rough–Smooth (R–S) and Hard–Soft (H–S) dimensions yielded moderate errors, performing better than S–S but not as reliably as F–B.

An analysis of modality contributions reveals complementary strengths between visual and tactile inputs. As shown in Table 3.5, visual features were particularly effective for predicting the F–B attribute and outperformed tactile features in this dimension. This may be attributed to the ability of image-based descriptors to capture explicit surface patterns, whereas tactile signals, particularly acceleration, may introduce high-frequency artifacts during exploration. These distortions can result in inconsistencies such as bouncing that affect prediction accuracy.

Texture class-specific results, visualized in Figure 3.9, indicate that paper and denim (jeans) categories accounted for the highest errors across multiple attributes. Within the paper class, the largest MAE was recorded for H–S (10.87) and R–S (9.57). This can be explained by the wide variation in sample properties, which include both plain and highly textured surfaces. Since predicted perceptual ratings are derived from user judgments, such discrepancies may stem from inconsistencies in human perception. Previous research [7] suggests that perceptual decisions can be influenced by prior expectations, leading participants to misjudge subtle differences based on experience rather than actual stimulus input.

A similar trend was observed in the jeans category, especially for T27 (a smoother variant), where intra-class variation likely contributed to increased error. The mesh category also exhibited higher error in the S–S attribute, with an MAE of 10.31. This may be due to recording noise introduced by hard plastic and metal meshes, whose rigid geometries can generate unstable contact transitions that degrade tactile signal quality.

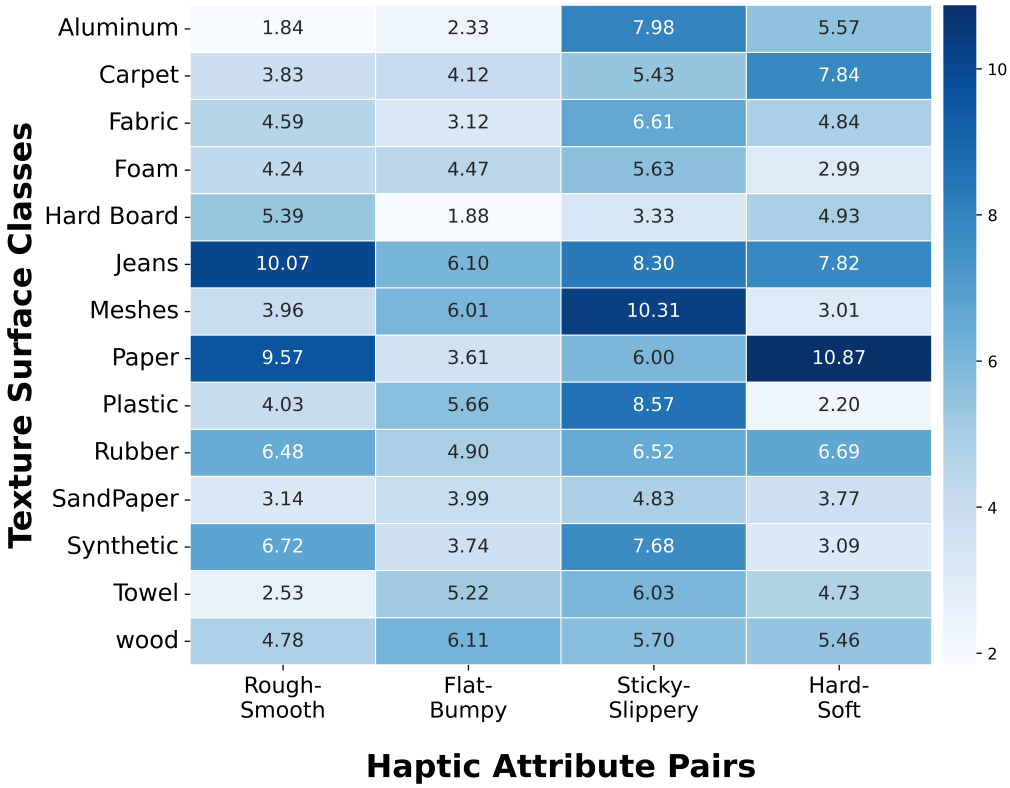


**Figure 3.9:** Class-wise Mean Absolute Error (MAE) distribution for haptic attribute prediction. The heatmap presents MAE values across texture classes and haptic attribute pairs, providing a detailed view of model performance. Darker regions correspond to higher prediction errors, while lighter regions indicate improved accuracy, revealing class-dependent variations in the visuo-tactile network's predictive capability.

Despite these localized discrepancies, the overall prediction errors remain within acceptable perceptual limits. Most of the class-wise and average MAE values fall below the Just Noticeable Difference (JND) threshold, commonly estimated at 10 on a normalized 0 to 100 scale [26]. This confirms the effectiveness of the proposed model in delivering perceptually aligned predictions across a broad range of texture types.

The generalization capability of the proposed model is further supported by its performance on distinct textures such as aluminum (T46), which presents unique surface characteristics. Despite its outlier status, the model performed reasonably well, with the highest error observed in the Sticky–Slippery (S–S) dimension at 7.98. This elevated error may be attributed to the presence of rubbing marks left by the interaction tool, a phenomenon previously noted in tactile research.

Since aluminum is the only sample within its category, additional analysis is warranted to understand this behavior in greater detail. Incorporating more samples with comparable surface properties is expected to enhance the model's robustness. At present, the study may not fully represent the spectrum of texture diversity required for comprehensive generalization.

Expanding the dataset with a broader range of texture types will likely improve prediction accuracy for haptic attributes. While Leave-One-Out Cross-Validation (LOOCV) can introduce certain biases, it remains a reliable strategy for evaluating model performance in small datasets. Overall, the results demonstrate that the proposed autoencoder-based architecture, which integrates CNN and handcrafted features, effectively captures subtle surface properties and outperforms prior unimodal baselines.

## 3.8 Conclusion

This study introduces a deep learning-based visuo-tactile framework specifically designed to estimate the perceptual attributes of textured surfaces. The model creates a mapping between a physical signal space, constituted by both visual and tactile features, and a perceptual space that is defined by user-assigned attributes. The perceptual space is structured according to four bipolar dimensions: rough-smooth, flat-bumpy, hard-soft, and sticky-slippery. The architecture incorporates a convolutional autoencoder for the encoding of visual features and a ConvLSTM network for capturing temporal patterns in tactile signals. Visual data is represented using features extracted from ResNet and GLCM, whereas tactile information is processed through MFCCs obtained from high-frequency acceleration recordings. The integration of these complementary modalities enhances the accuracy of predictions and supports improved generalization compared to existing unimodal approaches. The findings confirm the reliability and effectiveness of the proposed framework in predicting haptic attributes from physical signals. This methodology offers significant practical utility in contexts where user ratings are challenging to obtain, such as in extensive material databases or automated systems. Additionally, it may prove advantageous in robotic perception applications, where precise surface characterization is critical for tasks involving interaction and manipulation [13].

# Chapter 4

# Haptic Texture Modeling and Rendering

While the previous chapter focused on predicting perceptual attributes from physical interaction signals, this chapter shifts focus toward synthesizing haptic feedback that can simulate those perceptual experiences. The ability to generate realistic haptic textures is essential for virtual and remote applications, including e-commerce, design, training, and gaming, where tactile information plays a central role in enhancing user interaction [89, 90]. Rendering perceptually aligned tactile signals requires accurate modeling of texture-induced vibrations based on user interactions such as scanning speed and contact force.

## 4.1 Motivation

Over the past decades, several modeling techniques have been proposed to simulate surface feedback, including geometry-based methods [91], stochastic models [92], and data-driven contact dynamics approaches [11]. Among these, data-driven methods have shown superior capability in reproducing realistic vibrations by recording acceleration signals during tool-based texture exploration and mapping them to interaction parameters [11, 40]. However, many of these techniques rely on segmenting signals into short stationary intervals using algorithms such as AutoPARM and RCP [93, 94], and modeling each segment with AR/ARMA or LPC techniques [36]. These approaches often require complex preprocessing, including segmentation, parameter interpolation, and frequency-domain alignment, which limits their scalability and real-time usability.

Recent advancements in deep learning have addressed some of these limitations. Neural network-based methods have shown potential to reduce the need for segmentation and improve reconstruction quality [39], while more complex models like CNN-BiLSTM architectures have further improved accuracy [40]. However, these models are computationally heavy and pose chal-

lenges for real-time rendering scenarios.

To address this, we propose a lightweight and efficient framework for haptic texture modeling called the Fourier-enhanced Transformer Encoder Network (FoTEN). FoTEN integrates a transformer-based encoder architecture [27] with a Fourier feature encoding block, enabling the model to extract both spectral and temporal features from uniformly segmented input signals. Unlike traditional segmentation that depends on signal stationarity, our approach employs fixed-size sliding windows, simplifying preprocessing while preserving signal continuity.

The transformer encoder efficiently captures long-range dependencies in time series data through its attention mechanism, while the Fourier encoder enhances spectral representation of the acceleration signal. Together, these components support accurate reconstruction of texture-induced vibrations in both time and frequency domains. Additionally, the model is trained using the Huber loss function, which offers robustness to noise and improves generalization in natural interaction conditions.

The proposed method is evaluated against several state-of-the-art techniques. Quantitative results show improved reconstruction accuracy in terms of both time-domain and spectral metrics. Furthermore, a user study confirms the perceptual effectiveness of the rendered feedback, supporting the viability of FoTEN as a real-time haptic synthesis model.

## 4.2 Overview

An overview of the overall system is shown in Fig. 4.1. The upper section displays the modeling approach; highlighting data collection/preprocessing (Sec.4.3), and model construction stages (Sec. 4.4). The lower part illustrates the rendering process (Sec. 4.5), where the haptic texture feedback is generated by a vibrotactile actuator using a deep learning model in a stylus-based application. This framework is further evaluated and analyzed numerically in Sec. 4.6, and perceptually in 4.7. Lastly, we discuss the overall framework in Sec. 4.8 and conclude our work in Sec. 4.9.

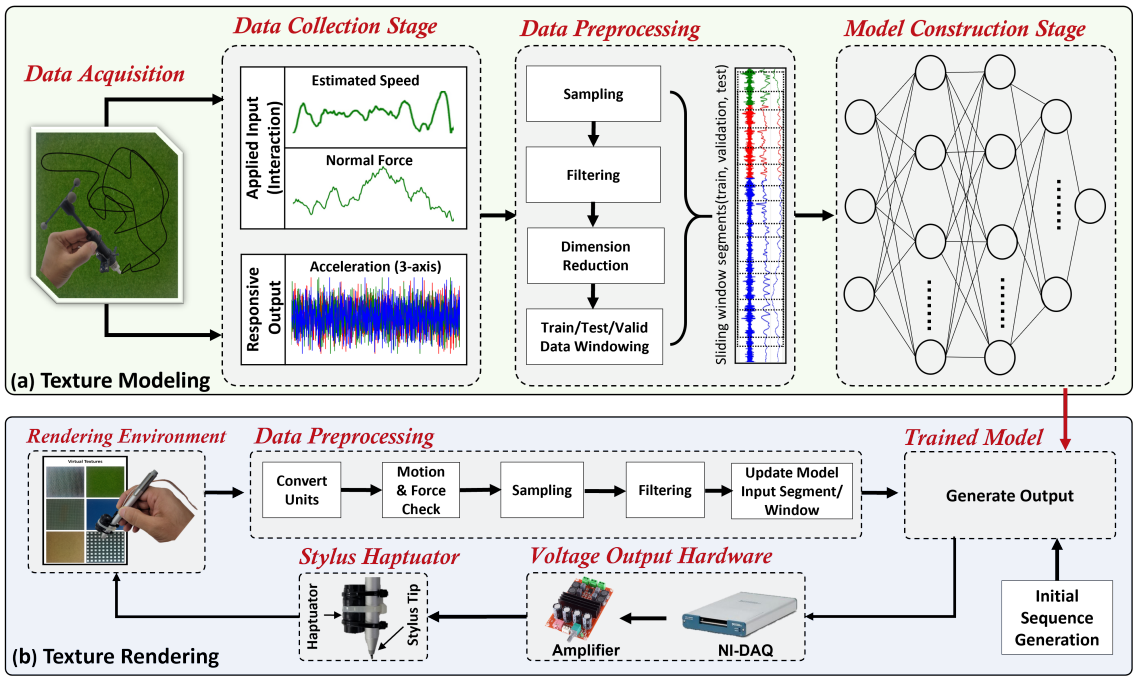


**Figure 4.1:** The overall framework. (a) Texture Modeling; The data acquisition setup is shown in the top left which is used for data collection. The vibration signal produced in response to the applied force and speed is recorded and processed in the next step. This processed is data then passed to the model construction stage. (b) Texture rendering; The below image illustrates the rendering of synthesized texture along with the hardware used for it.

## 4.3 Data Acquisition

Developing a haptic texture model starts with the acquisition of data while interacting with a textured surface using a sensorized tool. This section initially outlines the hardware setup and the texture samples used in this study. Then we discuss the processes of data recording and the essential pre-processing steps.

### 4.3.1 Hardware Setup

The data acquisition system used in this study can be seen in Fig. 4.2, utilizes a custom rigid tool equipped with an interchangeable tip for interacting with textured surfaces. The body of the rigid tool is custom-designed and fabricated using a 3D printer with ABS-Plastic material, while the attached hemispherical tool-tip is made of stainless steel and has a 2.0 mm diameter. A 3-axis

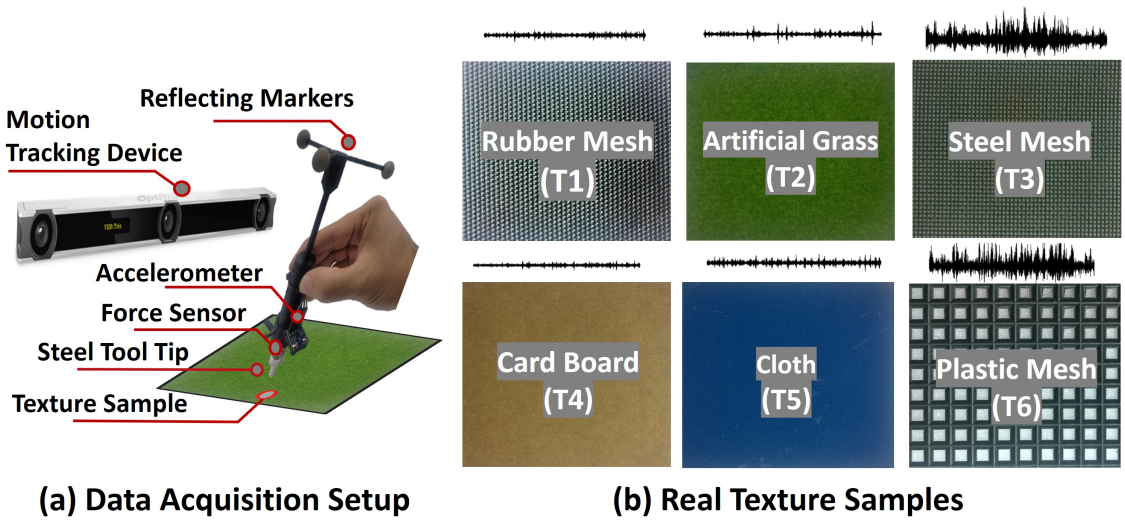


**Figure 4.2:** Hardware setup designed for capturing 3-axis vibrations elicited from textured surfaces, tracking interaction motion, and measuring applied force.

accelerometer (ADXL335; Analog Devices) was tightly attached to record the induced vibrations accurately while reflective markers were positioned at the upper end of the tool to track motion by an external position tracker (Optitrack: V120). Additionally, it incorporates a force sensor (Nano17; ATI Industrial Automation) to measure 3-axis interaction forces. Both the force sensor and accelerometer are connected to the PC through a data acquisition card (USB-6351; National Instrument).

### 4.3.2 Texture Samples

The selection of texture samples was made carefully to encompass a wide range of physical properties, including slipperiness, fineness, hardness, roughness, and bumpiness. This deliberate choice ensures that, despite the concise nature of the dataset, it remains representative of a broad spectrum of texture categories and effectively demonstrates the applicability of the proposed framework. Accordingly, six isotropic texture samples were chosen for performance evaluation: textured rubber, artificial grass, steel mesh, cloth, cardboard, and plastic mesh, as depicted in Fig.4.2. Each of these texture samples was cut into planar 100 mm square sheets and mounted on hard acrylic surfaces measuring 100x100x5 mm using liquid surface glue. This method of mounting helps avoid the influence of underlying objects during data recording and perceptual experiments.

### 4.3.3 Data Collection and Pre-processing

In our study, we prepared 6 different texture samples for data recording, as depicted in Fig.4.2. For each texture, data was collected for 20 seconds by unconstrained manual stroking of the rigid tool against the textured surfaces. The 3D position data is captured at 120 Hz and projected onto the tangential plane of the contacting surface. Subsequently, this position is used to estimate the scanning speed. The applied 3D-Force data is sampled at 10 KHz and projected onto the normal direction of the contacting surface to estimate the scalar normal force. Acceleration data from the 3-axis accelerometer is captured at 1000 Hz. All these signals are up/down-sampled at 1000Hz for synchronization.

Recorded interaction signals (i.e., scanning speed and normal force) are low-pass-filtered at 25 Hz, whereas the recorded acceleration signals are band-passed filtered from 20Hz to 1000Hz. The purpose of these filters is to suppress unnecessary noise and to remove the gravitational component. Next, these 3-axis acceleration signals are mapped onto a single axis using the DFT321 algorithm, which can capture both the temporal information and spectral energy of all three axes [95]. Lastly, we split and normalized the three signals (scanning speed, normal force, and acceleration) into the train ($70\%$) validation ($20\%$) and test ($10\%$) parts. For normalization, the training set details were utilized, and these parameters were preserved for rendering.

## 4.4 Modeling Approach (Fourier Enhanced Transformer Encoder Network)

The Transformer network, often referred to as the 'Attention-is-All-You-Need' architecture, was pioneered by Vaswani et al. [27]. It typically consists of an equal number of layers in both the encoder and decoder blocks, usually six each. However, depending on the specific application or modifications to the architecture, the number of layers in the encoder and decoder can vary independently while ensuring the internal structure of the encoder/decoder layer is consistent, to get the most out of the transformer's core mechanism: the multi-head attention mechanism. For instance, it is significant to use the encoder-decoder network in seq2seq tasks which involves converting one sequence to another (e.g., machine translation, question-answering, text summarizing,

etc ) where the decoder is generally used for generative purposes. On the other hand, recently, encoder-only networks are adept where tasks involve a single sequence as input for generating a high-dimensional representation (feature extraction) of that input for various applications, i.e., time-series forecasting, ECG classification, and regression.

In light of the aforementioned work's success, in this paper, we design a transformer encoder-only architecture to predict the acceleration signal based on historical data (i.e., predefined length of previous acceleration signal, interaction speed, and force.). To the best of our knowledge, this is the first attempt to model and render the haptic texture models using the transformer encoder layer-based architecture. The encoder-only structure is preferred over the encoder-decoder architecture while considering efficient processing and fewer parameters. Moreover, since the input and output lengths are predefined and fixed, the decoder's generative capabilities are not essential. Nevertheless, our contribution also stands out in the inclusion of spectral features in our network, aiming to improve the quality of the synthesized signal not just in the time domain but in the spectral domain as well. We explored various spectral features well-documented in haptic modeling literature, including Mel Frequency Cepstral Coefficients (MFCC) [15], wavelet transform [96], and Fourier transform [39]. Through rigorous evaluation, we determined that the Fourier transform when combined with the transformer encoder, yields the most promising results. These layers enable the network to recognize patterns and variations in signal frequencies. Thus, it is beneficial in capturing the global context of the entire signal in the spectral domain. The architecture of the proposed FoTEN can be seen in Fig. 4.3. It is composed of two parallel independent blocks namely the Encoder block and Fourier-encoder block. Below, the model's input and detailed explanations of each block are provided.

### 4.4.1 Model Input

The proposed network is designed to predict the acceleration $a[n]$ at time $n$ by taking a fixed-length $m$ sequence of previous acceleration $a$, scanning speed $v$, and applied force $f$ as input. The input for each time point $n$ is derived using a single-step sliding window mechanism, which can be represented as :

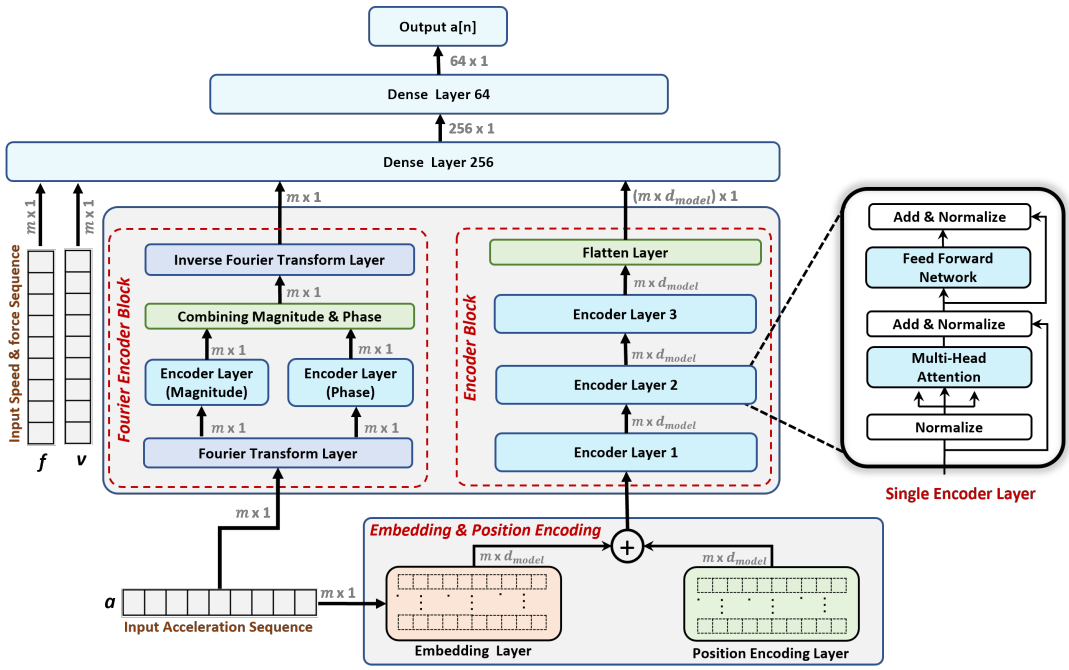


**Figure 4.3:** Structure of the proposed Fourier Enhanced Transformer Encoder Network (FoTEN).

$$\begin{aligned} \text{Input}_{\text{Sequence}}[n] = (a_{n-m}, \ldots, a_{n-1}, \quad & v_{n-m}, \ldots, v_{n-1}, \\ & f_{n-m}, \ldots, f_{n-1}), \end{aligned} \tag{4.1}$$

where $m$ represents the length of the sliding window. This approach allows the model to consistently use recent data for predictions and effectively adapts to data shifts or trends in the data over time.

### 4.4.2 Position Encoding

Position encoding is a crucial component for non-recurrent models like transformers to preserve the temporal sequence or the order of input elements. Unlike recurrent models such as RNNs or LSTMs which process data sequentially and inherently understand the order of input elements, transformers process data in parallel. This parallel processing significantly reduces the training and inference time and allows for greater scalability. However, it leaves the network without an inherent trace of the positional order of elements within the given input sequence, necessitating the use of position encoding to provide this critical information.

The original transformer paper [27] proposes a solution by integrating position encoding with embedding layers. Embedding layers are typically dense layers whose primary goal is to project the input feature vector into a higher-dimensional, learnable space to extract more complex features. Position encoding, on the other hand, is a set of distinct vectors, typically generated using sine and cosine functions. The use of sine and cosine functions for position encoding is crucial, as it allows models to distinguish sequence positions more effectively than simple numerical increments. This approach avoids the risk of disproportionately scaling input features—a common issue with integer-based encoding. By maintaining positional information in a scale-invariant manner, sine and cosine functions ensure that the model can interpret positional data without altering the input feature's magnitude undesirably. The positional encoding vectors are generated using sine and cosine functions for even and odd dimensions, respectively:

For even indices:

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right) \tag{4.2}$$

For odd indices:

$$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right) \tag{4.3}$$

where, $pos$ represents the position within the sequence, and $i$ is the dimension index, with $d_{\text{model}}$ indicating the size of the embedding space. This approach ensures each position in the sequence is encoded with a unique pattern by alternating between sine for even and cosine for odd dimensions. These are then added to the embedding space before being forwarded to the transformer encoder block, which has a size of $m \times d_{\text{model}}$, where $m$ represents the length of the input sequence.

### 4.4.3 Encoder block

The encoder block of our architecture comprises three standard Transformer encoder layers, identical in structure to the model proposed by Vaswani et al. [27]. The input to this encoder block is a segment of the acceleration signal, which is transformed and augmented with embeddings and positional encoding, resulting in dimensions of $m$ x $d_{\text{model}}$ where $m$ represents the sequence length and $d_{\text{model}}$ denotes the dimensions of the embedding layer.

Initially, these transformed acceleration signals undergo normalization as the first step in the encoder layer. These normalized signals are then further passed to a multi-head self-attention mechanism to capture spatial and temporal relationships between varying time steps in the given input. The multi-head attention mechanism operates with several scaled dot-product attention units in parallel, referred to as 'heads.' This configuration allows each head or attention mechanism to focus on different parts/sub-spaces of the input sequence independently by generating unique "attention weights." Each attention transforms the given input $X$, into queries $Q = XW^Q$, keys $K = XW^K$, and values $V = XW^V$, using learned weight matrices $W^Q$, $W^K$, and $W^V$ for Q, K, and V, respectively. Queries allow the model to focus on current points of interest, keys compare these points to others, and values prioritize content based on these comparisons (values), thereby enhancing its pattern recognition and generation capabilities. To integrate multi-head attention and to attend to different parts of the sequence simultaneously during feature extraction, the transformer architecture combines the outputs from individual heads as follows:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O, \tag{4.4}$$

where each head $i$ computes:

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V). \tag{4.5}$$

$W^O$ is a learned weight matrix that integrates the attention results from all heads into a cohesive output. This multi-head architecture allows the model to capture a richer representation of the input data. After the attention layer, the encoder layer employs residual connections and normalization to refine the output, which is then processed by a Feed-Forward Network (FFN) with a hidden layer of 128 units. Again, the output from the Feed-Forward Network is added back to its input (another residual connection) and normalized. It is to be noted that the input and output sizes of the encoder layer were kept the same to further pass to the remaining encoder layer to repeat the process of rich feature extraction in a black-box manner.

In our exhaustive evaluation, factoring in the number of heads and encoder layers as adjustable hyper-parameters, we determined that a configuration of three encoder layers, each with four

heads, achieves superior feature extraction to predict the precise acceleration signals.

### 4.4.4 Fourier encoder block

In the Fourier encoder block of our architecture, the input acceleration signal $a(n)$, with dimensions $m \times 1$, undergoes a transformation through the Fast Fourier Transform (FFT) that converts it from the time domain to the frequency domain:

$$Z(k) = \sum_{n=0}^{m-1} a(n) \cdot e^{-i2\pi kn/m} \tag{4.6}$$

In this equation, $a(n)$ is the acceleration of the signal at time index $n$, $m$ is the total number of samples, $k$ is the frequency index, and $Z(k)$ is the complex number representing the frequency component at $k$.

This transformation yields $m$ complex numbers, each inherently linked to a specific frequency. Unlike time-domain sequences where the order of samples dictates their interpretation, the components in the frequency domain are naturally ordered by frequency. Each frequency $k$ serves as a unique identifier, and its position within the sequence provides all necessary information regarding its characteristics in the signal, making additional positional encoding unnecessary.

After the FFT, the complex numbers are separated into their magnitude $M(k) = |Z(k)|$ and phase $\phi(k) = \arg(Z(k))$ components, both maintaining the $m \times 1$ format. These components are processed in dedicated encoder layers for magnitude and phase. This specialized processing allows the model to distinctly focus on the amplitude variations and the timing relationships among frequencies, enhancing the representation of the signal's frequency domain characteristics.

The encoded magnitude and phase outputs, $\hat{M}(k)$ and $\hat{\phi}(k)$, are subsequently recombined:

$$\hat{z}(k) = \hat{M}(k) \cdot e^{i\hat{\phi}(k)} \tag{4.7}$$

These recombined complex numbers, still in an array of $m$ components, are then transformed back to the time domain using an Inverse Fourier Transform (IFT):

$$f(n) = \frac{1}{m} \sum_{k=0}^{m-1} \hat{z}(k) \cdot e^{i2\pi kn/m} \tag{4.8}$$

This transformation, from time domain to frequency domain and back, with encoding and decoding in between, ensures that the output, also $m \times 1$, preserves the enhanced characteristics learned from the frequency domain. By leveraging the Fourier transform, the block effectively captures and encodes long-term dependencies and complex frequency-related features of the signal. The omission of positional encoding is deliberate; in the frequency domain, the inherent properties of the data ensure that each component's significance is maintained without additional encoding, facilitating a deeper understanding and more accurate prediction of the signal's behavior over time.

### 4.4.5 Network Training

The FoTEN model is implemented and trained as a unified model using the Python Keras-TensorFlow 2.8 library. The input acceleration signal, with dimensions $m \times 1$, is passed to the Fourier encoder block without positional encoding, while it undergoes positional encoding when passed to the traditional encoder block. In the Fourier encoder block, the signal is transformed to the frequency domain using a Fast Fourier Transform (FFT), resulting in $m$ complex numbers. These complex numbers are separated into magnitude and phase components, which are processed through dedicated encoder layers. The output of the Fourier encoder block, with dimensions $m \times 1$, and the output of the traditional encoder block, with dimensions $m \times d_{\text{model}} \times 1$, are concatenated along with additional interaction parameters: speed ($v$) and applied force ($f$), both of size $m \times 1$. This concatenated feature vector is then passed through two dense layers with 256 and 64 units, respectively, followed by a final regression layer that produces the acceleration output of size $1 \times 1$.

Given the complexity of the time-series data and the sophisticated model architecture, two critical hyper-parameters require careful consideration: the length of the input sequence and the choice of the loss function. These factors significantly influence the model's capability to capture and learn meaningful patterns from the data, as suggested in various studies [40, 97]. We hypothesize that optimizing these parameters will enhance the model's performance.

**Input Sequence Size** The length of the temporal window used in input data segmentation is a critical hyper-parameter that significantly impacts the model's ability to capture both spatial

and temporal patterns. Each input sequence is treated as an independent signal by the network, and this approach is also used for data augmentation. We adopted a sliding window-based data augmentation strategy [98], which enhances the diversity of the training data and improves the model's robustness. To determine the optimal window size ($m$ in eq. 4.1 or the input sequence size), we evaluated five different window sizes: 10, 14, 20, 26, and 30 samples. This comparative analysis aims to identify the window size that best captures the relevant patterns in the acceleration signal. We hypothesize that an optimal window size exists that balances the trade-off between capturing sufficient temporal context and maintaining computational efficiency.

**Loss Function** The choice of loss function is another crucial aspect of the model's design, especially when dealing with complex time-series data that may contain outliers and irregularities. Loss functions play a pivotal role in training the model by computing the error between the predicted outputs and the actual target values. During backpropagation, the computed loss is used to update the model's weights, guiding the learning process to minimize this error. To address this, we experimented with three different loss functions using the Python Keras-TensorFlow library: Mean Absolute Error (MAE), Mean Squared Error (MSE), and the Huber Loss (HL) function. MAE is less sensitive to outliers, treating all errors equally, whereas MSE gives more weight to larger errors due to the squaring of discrepancies, making it more sensitive to significant errors. The HL function is defined as:

$$L_\delta(a) = \begin{cases} \frac{1}{2}a^2 & \text{if } |a| \leq \delta, \\ \delta|a| - \frac{1}{2}\delta^2 & \text{otherwise,} \end{cases} \tag{4.9}$$

where $L_\delta(a)$ is the Huber loss, $a$ denotes the error between true and predicted values, and $\delta$ is a threshold value, that combines the robustness of both MAE and MSE. For errors less than $\delta$, the HL behaves like MSE, adopting a quadratic nature, while for larger errors, it behaves like MAE, becoming linear. This property enables HL to be less sensitive to outliers compared to MSE. We determined the optimal value of $\delta$ through experimentation, assessing model performance using $\delta$ = 1.2, 1.6, and 2.0. We hypothesize that the Huber Loss function, with its combined properties, will provide a balance between sensitivity to outliers and the ability to penalize large errors, thus improving overall model robustness [97].

For training, the model uses the ADAM optimizer with a learning rate of 0.001 and a batch size of 16. Training is carried out for a maximum of 100 epochs; however, an early stopping mechanism is incorporated to stop training if there is no improvement in the validation loss for 10 consecutive epochs, preventing over-fitting. Evaluation metrics, including mean absolute error (MAE), mean squared error (MSE), and Huber loss, are employed and will be further detailed in a subsequent section. Moreover, within the entire network, the RELU function acts as the activation function while a dropout rate of 0.2 is employed as a regularization technique.

## 4.5 Texture Rendering

This section outlines the procedure of synthesizing textures using the transformer-based models established in Sec. 4.4. Initially, we discuss the rendering hardware setup used in this study. Later, the detail of the crafted approach is described to generate vibration signals for user feedback.

### 4.5.1 Signal Synthesis

The goal of the rendering algorithm is to haptically synthesize a series of acceleration sequences according to the user's interaction, i.e., stroking the pen on the tablet PC's screen. More specifically, the algorithm produces vibration output $a[n]$ at time $n$ by observing the three input data; previous sequence of acceleration $a$, user's stroking speed $v$, and pusing force $f$, each of which has size of $m$ as detailed in eq. 4.1.

At the initial moment of contact, we do not have initial sequence of data for model's input of size m (e.g., 20 ms in case of m = 20). So we need a way of estimating these initial input data for acceleration $a$, user's stroking speed $v$, and force $f$. For $v$ and $f$ we can begin capturing data as soon as the user initiates interaction, ensuring that our inputs reflect real user behavior. However, for initial acceleration the estimation of the sequence is needed. This is crucial since large discrepancies may lead to slow convergence of the model estimation depending upon the generalizability of the model [99]. In literature, such initial acceleration sequence was mainly generated by randomization [39] and through weighted interpolation of stored segments [40]. However, these techniques considered best suited for their approach and were contextually adapted. As of

our observations, random initialization could take longer for the model to converge while interpolation of the stored segments requires various stationary segments to be loaded in the memory. Furthermore, such interpolation may increase inference time, high memory usage and potentially causes artifacts in real time rendering as reported by [40].

In contrast, we employed a hybrid feature based technique consisting of offline computing of acceleration signal using our trained model. The main goal of this is to estimate the acceleration signal for the interaction speed and force which user likely to interact with initially. Thus, under the assumption that user interaction initially involves very small amounts of speed $v$ and force $f$, as well as minimal variations in these parameters. This was also confirmed when investigated with the interaction data recorded for model training. For the first second of interaction across all six textures, the mean speed $v$ was found approximately 45 m/s and the force as 0.8 N. Considering these parameters as representations of the user's initial interaction, we synthesized an acceleration signal of length $m$ at 45 m/s and 0.8 N, which we stored for model input. This initial sequence provides two core benefits: it does not require the storage of various stationary acceleration segments, and the acceleration sequence generated defines the initial interaction more strategically rather than relying on random initialization or a feature-based strategy and hence resulting in efficient estimation with good performance (see Sec. 4.7).

Finally, the initial input to the FoTEN is precomputed by the estimation of an acceleration signal from our trained model and is used along with initial interaction data of $v$ and $f$. Once the initial input interaction parameter sequence is completed the FoTEN model get its first input. In subsequent steps, the acceleration sequence is updated with FoTEN output while $v$ and $f$ update according to the user's interaction, maintaining the recent values. This process occurs within a moving window, ensuring continuous updating and incorporation of the latest data.

### 4.5.2 Rendering Hardware

The complete illustration of the texture rendering hardware setup is depicted in Fig. 4.1(b), which we utilized for the psychophysical study detailed in Sect. 4.7. A touch tablet PC (Surface Pro 4; Microsoft) equipped with an active digital stylus (Surface Pen; Microsoft) is selected as the interface for rendering. The vibration feedback for the virtual texture was generated using a high-

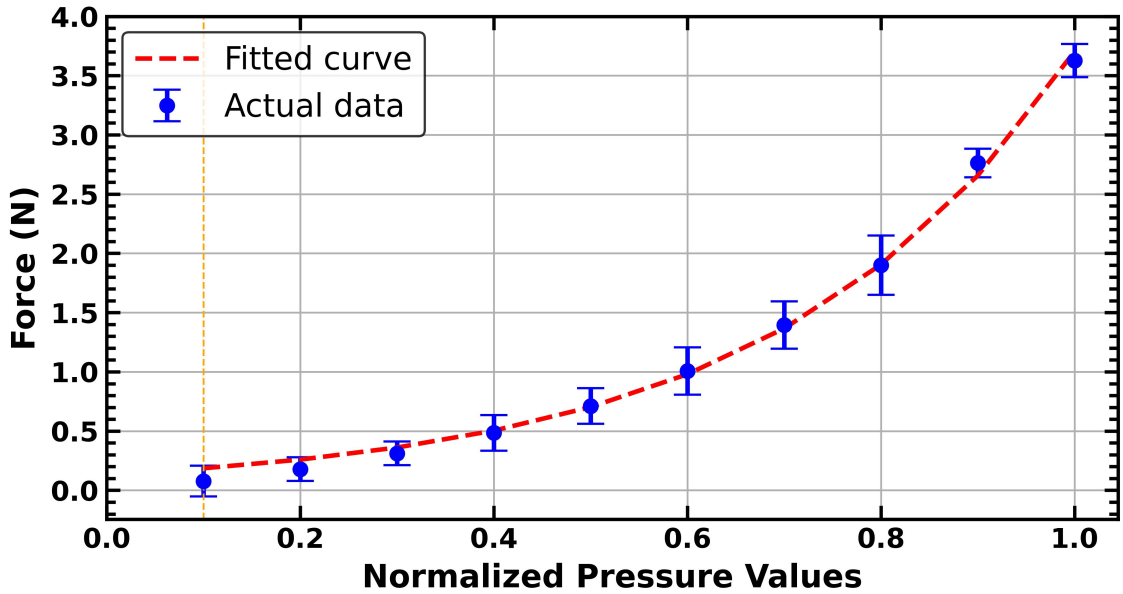


**Figure 4.4:** Relationship between normalized pressure values and recorded force magnitude for the tablet.

bandwidth voice-coil actuator (Haptuator MM1C; Tactile Labs) attached to the stylus pen near the tool tip with the help of plastic zip-ties. The haptuator is controlled via the NI-DAQ device (USB-6351; National Instruments) to render textures at 1000Hz while an analog amplifier is also employed to control the gain of rendered signal and acts as a bridge between NI-DAQ and the haptutor. Moreover, a feed-forward dynamic compensation controller was also utilized to counter the haptuator's natural frequency response dynamics similar to [100].

For our model input, the magnitude of applied force was computed from the touch stylus, which detects 1024 pressure levels upon contact with the screen. However, the PyQT library provides normalized values within a range from 0 to 1 and our model requires force in standardized unit Newtons (N). To model the applied force accurately, an experiment was conducted to establish a quantitative relationship between the pressure levels and their corresponding force values. For this, we positioned a force sensor under the tablet, applied manual pressure in increments of 0.1 to the screen's normal direction, and recorded the resulting forces. The collected data was then analyzed using the Scipy Library, which helped us identify an optimal exponential model. This model was chosen based on the observation that the force response from the stylus exhibits a rapid increase with slight increments in pressure, a behavior not as effectively captured by linear or

polynomial models. The relationship is encapsulated by the following equation:

$$\mathrm{F} = 0.13 \cdot \exp(3.32 \cdot y) \tag{4.10}$$

where $F$ is the force in Newtons and $y$ is the normalized pressure value obtained from the software.

### 4.5.3 Rendering Software

A user interface has been developed to render texture. This rendering application is designed to display a texture image of the same size as our actual texture sample. The software detects three inputs upon interaction: the texture type to load the initial sequence (see Sect. 4.5.1), the sliding speed of the stylus, and the applied pressure. The speed $v$ calculations are based on the $x$ and $y$ positions captured from the screen during interactions within the 2D space, specifically when the user interacts with the area displaying the texture. Meanwhile, the force is calculated using pressure values from Eq.4.10.

Thresholding is implemented to enhance the user experience upon releasing the stylus or when the stylus is not moving. For this purpose, the minimum speed threshold is set to be greater than 10 mm/s, and the force to be more than 0.15 N. Additionally, the maximum threshold for speed was set at 300 mm/s and for force at 3.6 N. When values fall below the minimum thresholds, it is assumed that the interaction has ceased, and all variables are set to zero, with the vibration output also set to 0. Conversely, when maximum threshold values are reached, we send the maximum value to the model input.

It is noted that the speed and force values are first upsampled to 1000 Hz, as the model requires, and then filtered at 20 Hz. This step ensures that the sampling rate matches that of the data used for model training and is filtered to remove several unwanted effects, such as noise, DC position offset, and quantization effects. These processed speed and force values are then used to update the input sequence as detailed in Sect. 4.5.1.

Finally, the acceleration output from the model is denormalized using pre-stored statistical parameters and sent to the NI DAQ after a dynamic compensation filter [100] to mitigate the haptuator's natural frequency response. This prepares it for vibrotactile feedback through a current amplifier.

## 4.6 Model Prediction Accuracy Measures

To evaluate the proposed model's performance, we collected tactile signals from six different textures (see Sec.4.3). Each texture data consisted of 20-second recordings at a 1kHz sampling rate, resulting in 20,000 samples per texture. The data was divided into 20 frames and regrouped based on speed and force regions, following the method in [90]. This regrouping ensures overlapping speed and force regions between training, validation, and test sets, maintaining a similar data distribution. The overall split ratio was set to training (70%), validation (20%), and test sets (10%) for each texture. The training and validation datasets were used for model training, while the test dataset, unseen during training, was used to evaluate the model's performance, including in ablation studies.

### 4.6.1 Error Metrics

Comparisons between synthesized acceleration signals and actual recorded acceleration waveform were performed in both the time and spectral domains to gauge the proposed model's performance. In the time domain, prediction accuracy was quantified using mean absolute error (MAE). For spectral domain comparison, we utilized the Hernandez-Andres Goodness-of-Fit Criterion (GFC) to measure the degree to which the reconstructed signal matches with the actual signal. The GFC score values range from 0 to 1, where a value of one is considered a perfect reconstruction [101]. Mathematically it can be computed as:

$$GFC = \frac{\left\| \sum_i A_d\left(f_i\right) A_m\left(f_i\right) \right\|}{\sqrt{\left\| \sum_j \left[A_d\left(f_j\right)\right]^2 \right\|} \sqrt{\left\| \sum_k \left[A_m\left(f_k\right)\right]^2 \right\|}}, \tag{4.11}$$

where $A_d(f_i)$ and $A_m(f_i)$ are the DFT amplitudes at a frequency $f_i$ of the measured and reconstructed signals, respectively. These metrics were chosen due to their frequent use by various authors in evaluating haptic texture modeling [74, 101].

### 4.6.2 Finding optimum sequence size and loss function

As part of our ablation study, we examine how input sequence size and loss functions affect model performance. Identifying the temporal length of the input sequence is crucial for accurate pre-

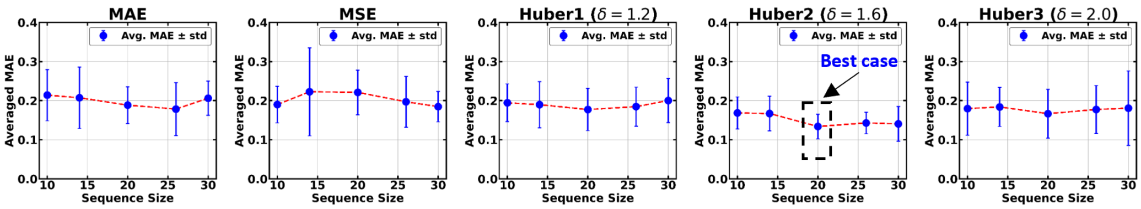


**Figure 4.5:** The comparison of model performance for various sequence sizes (i.e., 10, 15, 20, 25, and 30 ) and loss functions (MAE, MSE, Huber) as part of our ablation study on the test dataset.

dictions [40, 102], and selecting an effective loss function is essential for improving training outcomes [97]. Importantly, loss functions aid in effectively updating the model's weights during training, while error metrics evaluate the performance of the trained model on test datasets.

For this study, we tested various input sequence lengths $m$ (see Eq.4.1) and loss functions, but the results focus on the most promising configurations, including five input lengths (i.e., $m = 10, 15, 20, 25, 30$) and three loss functions: MAE, MSE, and Huber loss, as identified in the literature [74, 97]. Huber loss (HL), which combines the effects of MAE and MSE with a controllable $\delta$ parameter, was further ablated with $\delta$ values of 1.2, 1.6, and 2.0. The results of these 25 experiments on test data are presented in Fig. 4.5 as averaged MAE. Observations indicate that HL ($\delta = 1.6$) with $m = 20$ outperformed MAE, MSE, and other sequence sizes by showing the lowest mean MAE of 0.148. Under these optimal conditions ( $\delta = 1.6$ and $m = 20$), the time-domain comparison of the synthesized and actual signals for the initial 400 samples is shown in Fig. 4.6. The Power Spectral Density (PSD) comparison, which effectively captures the frequency domain details, is illustrated in Fig. 4.7 for the entire test signals for all 6 textures, demonstrating a reconstruction that is challenging to accomplish.

### 4.6.3 Comparison with other approaches and spectral features

The signal reconstruction accuracy of the proposed framework (FoTEN) was also evaluated against various established methods. These methods include traditional AR based technique [11], a neural network strategy (FNN) [39], and advanced deep learning based spatio temporal network (DSTN) [40], GAN based technique [38], and the Haptic Informer network, which employs a transformer-based encoder-decoder architecture [74]. Additionally, we also assessed our architecture's variants: with and without the Fourier/Spectral block (TEN) and with texture features

**Table 4.1:** Comparison of error metrics (MAE and GFC) across existing methods and textures, with modeling types and features.

| Study | Modeling Type | Segmentation Technique | Modeling Features | Metric | Textures | | | | | | Mean |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | T1 | T2 | T3 | T4 | T5 | T6 | |
| **Piece-wise AR** [11] (2014) | Stochastic | AutoParm | Raw Acceleration | MAE | 0.48 | 0.36 | 0.39 | 0.41 | 0.27 | 0.53 | 0.406 |
| | | | | GFC | 87.56 | 80.58 | 86.58 | 85.16 | 91.18 | 88.79 | 86.64% |
| **FNN** [39] (2015) | NN | Constant Speed & Force | Raw Acceleration + Frequency Decomposition | MAE | 0.41 | 0.28 | 0.46 | 0.29 | 0.39 | 0.47 | 0.384 |
| | | | | GFC | 89.31 | 90.88 | 86.37 | 86.14 | 85.08 | 89.92 | 87.94% |
| **GAN-based** [38] (2018) | DL | Constant Speed & Force | Raw Acceleration + Spectrogram | MAE | 0.23 | 0.28 | 0.27 | 0.29 | 0.32 | 0.37 | 0.292 |
| | | | | GFC | 90.85 | 89.39 | 88.50 | 90.54 | 83.21 | 90.71 | 88.86% |
| **DSTN** [40] (2021) | DL | Constant Speed & Force | Raw Acceleration | MAE | 0.15 | 0.18 | 0.27 | 0.17 | 0.15 | 0.27 | 0.195 |
| | | | | GFC | 91.83 | 89.96 | 90.54 | 88.18 | 90.22 | 90.71 | 90.24% |
| **HapInf** [74] (2023) | DL | Sliding Window | Raw Acceleration | MAE | 0.18 | 0.16 | **0.12** | 0.15 | 0.17 | 0.29 | 0.176 |
| | | | | GFC | 89.14 | 92.02 | **91.48** | 92.01 | 91.36 | 89.38 | 90.89% |
| **FoTEN** (ours) | DL | Sliding Window | Raw Acceleration + Fourier Transform | MAE | **0.13** | **0.11** | 0.16 | **0.13** | **0.14** | **0.19** | **0.148** |
| | | | | GFC | **90.83** | **93.82** | 91.17 | **93.83** | **92.17** | **91.84** | **92.28%** |

including Mel-Frequency Cepstral Coefficient (MFCC) and Wavelet decomposition. For the TEN variant, we removed the Fourier/Spectral block; for MFCC and Wavelet, we adjusted the spectral block while keeping other components unchanged. These variations were further ablated for input window size and loss functions, similar to FoTEN.

The comparison results with other approaches for each texture, along with the experimental settings, are summarized in Table 4.1. FoTEN achieves the lowest mean MAE of 0.148 and the highest mean GFC of $92.28\%$, showcasing its effectiveness in both time and spectral domains. In contrast, the piece-wise AR and FNN showed less favorable outcomes in the time and frequency domains, respectively. HapInf ($90.89\%$) and DSTN ($90.24\%$) achieved close reconstruction accuracy in the spectral domain, while the GAN-based approach lagged behind all other DL methods. Notably, FoTEN surpassed most methods across different textures, except for Steel Mesh (T3), where the HapInf model recorded slightly better accuracy.

Table 4.2 displays the performance of our architecture with different features, including the input size on which each variant produced the lowest error. The analysis shows that using FFT as a spectral feature not only boosted our model's performance compared to other features but also reduced the input size. It is worth noting that the Wavelet variant produced these results with the MSE loss function, while MFCC and FFT performed better with HL($\delta = 1.6$) and TEN with HL($\delta = 2.0$). This analysis confirms that Fourier enhancement provides more robust results, often surpassing existing techniques with superior signal reconstruction capabilities, particularly with

the inclusion of Huber Loss, which is being investigated in texture modeling for the first time.

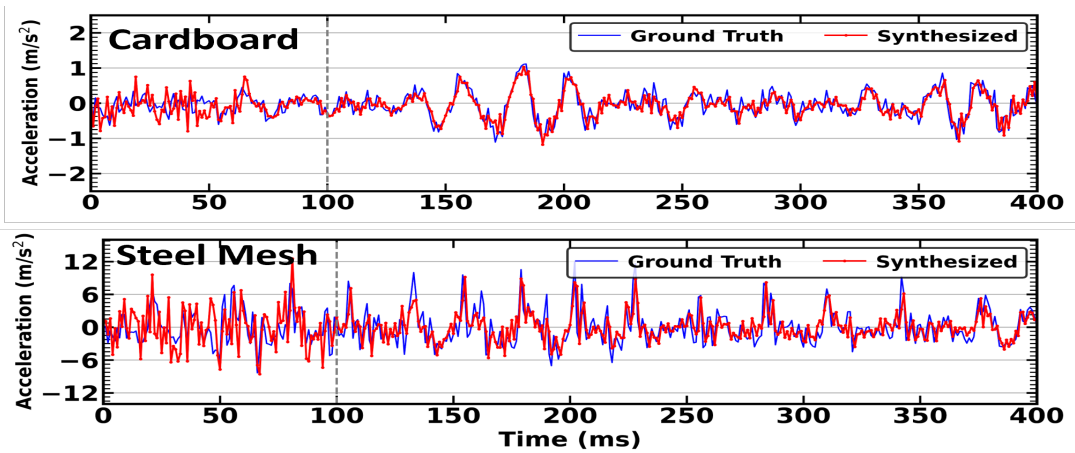


**Figure 4.6:** Time-domain comparative analysis of synthesized and recorded acceleration signals.

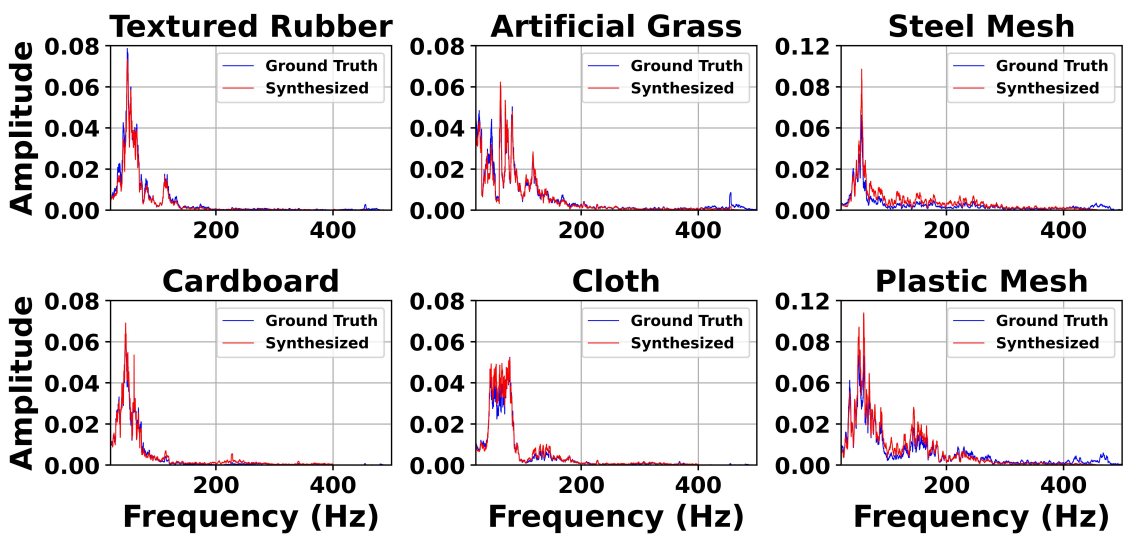


**Figure 4.7:** Spectral-domain analysis comparing recorded and synthesized acceleration signals for all 6 textures on test data.

### 4.6.4 Computational Efficiency

An important aspect of deep learning-based real-time rendering is model's prediction (inference) time. The FoTEN framework exhibits significantly lower inference times than previous methods like DSTN [40], GAN-based [38] and HapInf [74]. This improvement is evident on both

**Table 4.2:** Comparison of spectral features on our model

| Texture Features | Input Size | Mean MAE | Mean GFC (%) |
|---|---|---|---|
| TEN (Encoder Block) | 25 | 0.325 | 86.84 |
| TEN + MFCC | 30 | 0.276 | 90.13 |
| TEN + Wavelet | 30 | 0.301 | 87.22 |
| **TEN + FFT (FoTEN)** | **20** | **0.148** | **92.28** |

CPU (using TensorFlow's TFLite extension) and GPU (using the .h5 model extension). The performance results, averaged over 1000 iterations of inference, along with training time per epoch and model input sequence size, are shown in Table 4.3). It can be seen that TEN, a subset of our model, achieves better inference performance by omitting the spectral block, which reduces computational load at the expense of a trade-off in accuracy (see Table 4.2).

## 4.7 Perceptual Performance

A total of twelve participants (9 males and 3 females) with an average age of 28.4 took part in this study. They compared virtual textures with their corresponding real counterparts and rated their similarity on a scale of 0 to 100 (100 being completely the same). Each participant rated all six textures using the AR model [11] as a traditional baseline method, alongside three advanced deep learning-based techniques: DSTN [40], Haptic HapInf [74], and FoTEN (our method). Consequently, users were presented with six virtual-real comparisons, repeated across the four methods, for their assessment.

### 4.7.1 Procedure

The complete setup can be seen in Fig. 4.8, where users were seated in front of a table playing white noise to minimize environment noise. The table was divided into two parts with a divider to avoid visual cues. Users faced a screen running a GUI to record responses on the left side while experiencing real/virtual textures on the right side through our customized tablet PC. The tablet screen was modified using styrofoam, which included two cutouts of identical size to facilitate the comparison between virtual and real textures; real on the right and virtual on the left.

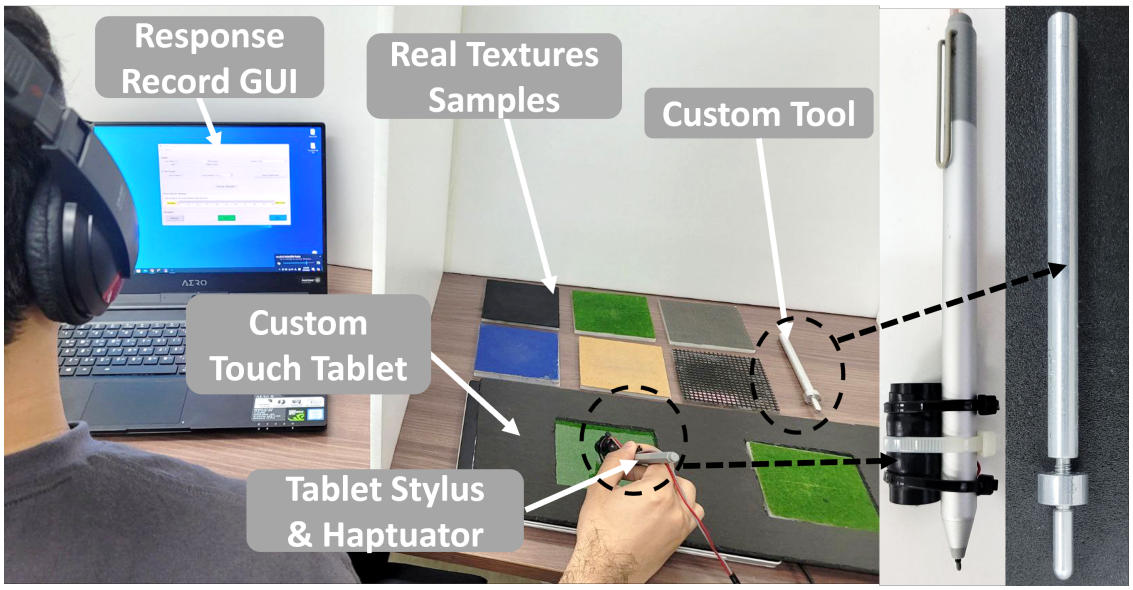


**Figure 4.8:** Experimental setup for the user study and the hand-held tools used in the experiments.

Users experienced virtual textures with stylus equipped with haptuator for vibration cues while for the real textures, a custom tool with the same tool-tip used for data recording was provided to maintain consistency in the modeled tactile feedback. They were asked to freely explore and switch between the real and virtual textures until they were satisfied. The experimenter assisted by guiding their hand and the tools while switching between textures. The order of the presented stimuli was randomized across textures, algorithms, and participants and was determined before the experiment. The experiment took an average of 55 minutes per user, and each was provided with 15,000KRW ($\sim$ 12 USD).

### 4.7.2 Results

The results of the perceptual experiment were in the form of similarity ratings ranging from 0 to 100. A total of 288 ratings were collected from 12 participants assessing four methods across six different materials. The results of the experiment are presented in the form of mean similarity in Fig.4.9. It can be seen that FoTEN received the highest similarity ratings across all real-virtual texture pairs among all methods, with steel mesh achieving the highest mean rating and cardboard the lowest.

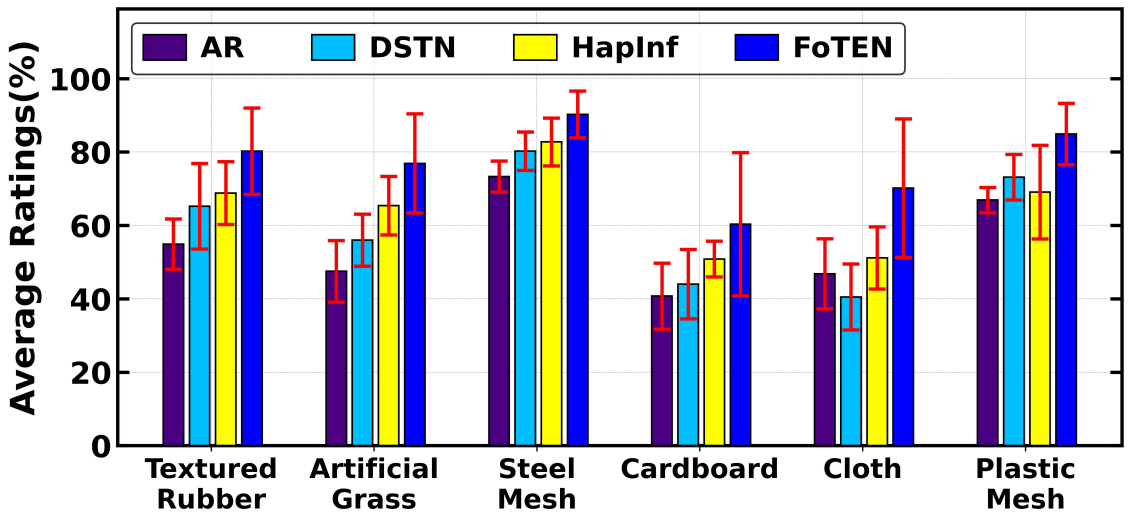


**Figure 4.9:** Perceptual ratings by textures and approaches.

A repeated-measures ANOVA was employed to evaluate the effectiveness of utilized method in capturing perceptual similarity, with the similarity rating as the dependent variable and methods as the within-subject factor. This approach was chosen because it allows analysis of the same subjects under multiple conditions, treating participants as a random effect to account for variability among them, thus enhancing the statistical power by focusing on method effects.

The ANOVA on the similarity ratings yielded a significant difference between methods, with a p-value of $<0.01$ ($F = 64.97$, $\eta^2 = 0.8077$), suggesting substantial effect size. This significant result prompted further investigation through pairwise comparisons, which were adjusted using the Bonferroni correction. The pairwise analysis revealed that FoTEN significantly outperformed AR Models, DSTN, and HapInf, with p-values of $<0.01$. The comparison between AR Models and HapInf also showed a significant difference ($p < 0.01$), highlighting notable disparities in their performance. In contrast, the comparisons between AR Models and DSTN, and DSTN and HapInf did not show significant differences, with p-values of $0.019$ and $0.138$ respectively, indicating closer performance metrics between these methods. The ANOVA on the similarity ratings yielded a significant difference between methods, with a p-value of $<0.01$ ($F = 64.97$, $\eta^2 = 0.8077$), indicating a substantial effect size. This significant result led to further investigation through adjusted pairwise comparisons.

## 4.8 Discussion

Both the numerical evaluation and perceptual experiments provided deep insights into the strengths and weaknesses of our modeling and rendering techniques. As shown in Table 4.1, the proposed approach, which utilized a simpler sliding window-based input, showed the highest signal reconstruction accuracy in both the time and spectral domains, outperforming existing stochastic and deep learning approaches by achieving a $92.28\%$ mean GFC score, and the lowest mean MAE of 0.148. This performance exceeded that of traditional approaches and was comparable with DL-based approaches. Moreover, since there is no predefined approach to initializing input sequences, we proposed an intuitive method that leverages the initial sequence by considering the user's natural initial interaction parameters.This contrasts with other approaches that rely on stored segments and frequent updates to initial segments [39,40], which can introduce artifacts in real-time rendering and affect realism. As illustrated in Fig. 4.6, our initial synthesis approach allows the model to achieve accurate predictions after a warm-up period of fewer than 100 iterations, while ensuring accuracy is maintained in both time and spectral domains throughout the prediction process.

Another key observation is that FoTEN requires less training and inference time compared to other DL-based approaches. In contrast, traditional methods [11] take significantly longer in the modeling process due to complex data preprocessing and intermediate steps, which demand consistent human-in-the-loop involvement from a haptic expert. While these methods are efficient in inference, this often comes at the expense of accuracy. Furthermore, FoTEN's faster performance compared to HapInf [74] highlights the effectiveness of modifications in the transformer architecture. It also outperforms BiLSTM-based networks [40], which are inherently slower due to their sequential nature. Conversely, the GAN-based approach [38], which generates spectrograms, is more computationally intensive and requires refinement for real-time applications.

The human-subject study demonstrated the model's perceptual performance, achieving an overall mean similarity rating of $77.14\% \pm 16.67\%$, and reported no artifacts with our approach. In contrast, DSTN exhibited artifacts, while HapInf showed delayed responses, as reported by users. Although perception tests showed that FoTEN effectively recreated the vibrations and roughness of real textures, it underperformed with finer and slipperier textures like cloth and cardboar. Post-experiment interviews suggested that materials with more pronounced textures, such as steel mesh

**Table 4.3:** Efficiency comparison of FoTEN and Existing Deep Learning based approaches

| Models | Input Seq. Size | Inference GPU(ms) | Inference CPU(ms) | Training One Epoch(s) |
|---|---|---|---|---|
| GAN [38] | 56 | 1.72 | 2.37 | 85.71 |
| DSTN [40] | 40 | 1.38 | 1.89 | 44.35 |
| HapInf [74] | 10 | 1.29 | 1.63 | 24.30 |
| **TEN(ours)** | 25 | 0.87 | 1.07 | 8.92 |
| **FoTEN(ours)** | 20 | 0.96 | 1.11 | 10.26 |

and textured rubber, felt more realistic. The main issue with finer textures likely stems from uncontrolled friction and stiffness in tool-based rendering of virtual textures, which complicates user's ability to distinguish between texture roughness and other characteristics, despite accurate signal reconstruction [36]. One solution for this is the consideration of other haptic properties such as softness and friction along with roughness while modeling and rendering, along with the utilization of force-feedback devices.

## 4.9 Conclusion

In this section, we presented a novel framework that combines Fourier transform with Transformer encoders to reproduce surface textures. We demonstrated that this method is lightweight and efficient for synthesizing real-time vibrotactile feedback. We simplified the modeling process by replacing complex data segmentation with a fixed-size sliding window approach and enhanced signal reconstruction accuracy by employing Huber loss during training. The data suggest that the proposed technique significantly enhances signal reconstruction accuracy in both the time and spectral domains, as well as perceptually, as validated by a user study, indicating substantial advancements over existing methods.

# Chapter 5

# Car Door Perception Modeling and Generation

Previous chapters addressed surface-level tactile perception, exploring how humans interpret textures through touch and how such impressions can be predicted or synthesized from visual-tactile signals. This chapter extends the scope of haptic modeling to kinesthetic feedback, using car doors as a representative example where user experience is shaped primarily by force and torque feedback. Unlike micro-level surface textures, car door interactions involve large-scale mechanical motion, governed by the door's mass, friction, and hinge dynamics.

## 5.1 Motivation

With the growing emphasis on autonomous driving and electric vehicle platforms, the focus of innovation in the automotive domain is shifting from mechanical performance to user-centered experience. Haptic interactions, in particular, are gaining attention as a key modality for shaping perceived quality and emotional response. Among these, the act of opening or closing a car door serves as a highly salient event. It is typically the first and last point of physical contact between the user and the vehicle, and strongly influences impressions of luxury, mechanical precision, and reliability [103–105].

Despite its importance, the process of tuning car door mechanics to elicit specific user impressions remains largely subjective. Automotive engineers often rely on physical prototyping and iterative adjustments, evaluating each design variation through hands-on testing and qualitative feedback [106]. While this method offers realism, it is time-consuming, expensive, and difficult to scale across multiple vehicle classes. As a result, perceptual tuning remains an informal practice rather than a systematic design task.

Virtual prototyping presents a promising alternative, allowing designers to simulate mechan-

ical interactions in a controlled environment without the need for repeated hardware revisions [107, 108]. In such systems, high-fidelity haptic feedback is essential to maintain realism and ensure that users can evaluate sensations with confidence [109]. Several studies have introduced task-specific haptic systems to simulate dials, refrigerator simulators, prosthetic hands, handles, and other physical interfaces [110–115], but few have systematically modeled how force profiles relate to user perception in car door interactions [108].

This study proposes a perception-centered, bidirectional modeling framework that learns the relationship between torque-based force signals and user-reported perceptual impressions. Rather than treating the interaction as a black box, the system defines a structured perceptual space using bipolar adjective pairs that reflect both physical and emotional qualities of door movement. These attributes are selected based on user input, literature review, and expert knowledge, ensuring that the space is both human-interpretable and relevant to design objectives [116].

By formulating the problem in terms of attribute-based modeling, the framework offers several advantages. First, it provides an explainable representation that helps researchers understand which signal features contribute to perceptual outcomes. Second, it enables controllable generation of force signals from intuitive input, allowing designers to specify desired experiences using meaningful terms such as ”comfortable” or ”recoiling” rather than abstract parameter values. Third, the perceptual space facilitates user studies, perceptual evaluations, and simulator-based prototyping with minimal ambiguity. Together, these qualities support the development of more transparent, adaptive, and user-aligned haptic systems.

In summary, this chapter addresses the challenge of modeling car door perception by combining real-world force measurements, psychophysical experiments, and deep learning architectures. The goal is to bridge the gap between mechanical signals and human perception using attribute-based modeling, enabling not only accurate prediction but also intentional generation of desired haptic experiences.

## 5.2 Overview

This study proposes a bidirectional learning framework that maps between the physical characteristics of a car door’s force profile and the perceptual impressions it elicits. The core objective is to

establish two structured spaces: a *Physical Signal Space* and a *Perceptual Attribute Space*, and to model both forward and reverse relationships between them. Figure 5.1 illustrates the full pipeline, including data collection, perceptual grounding, augmentation, and bidirectional learning.

### 5.2.1 Defining the Two Spaces

The *Physical Signal Space* consists of force profiles that describe how much effort is required to open a car door over time or angular displacement. These signals encapsulate mechanical properties such as mass distribution, hinge resistance, and damping effects. The *Perceptual Attribute Space* contains user-reported evaluations of these signals in terms of adjective-based cognitive impressions (such as smooth, heavy, or luxurious), collected through structured user studies.

Both spaces are built progressively. The physical signals originate from real car door recordings, while the perceptual attributes are grounded through multiple sources including literature, expert knowledge, and user feedback from interactions with real cars and a car door simulator.

### 5.2.2 From Data Collection to Augmentation

Force profiles were first recorded from six commercially available car doors using a torque sensor and optical tracker. These signals were preprocessed and normalized to a consistent angular domain. To allow repeatable, isolated haptic interaction, a car door simulator was developed to replay these signals physically using a programmable servo mechanism.

In parallel, perceptual adjectives were gathered through freeform and structured user studies conducted with real car interactions. Users were asked to describe their impressions after opening the doors, and these were combined with expert and literature-based sources. A second study was then conducted using the simulator to refine and shortlist the most meaningful attributes. Users highlighted the importance of initial motion phases, which informed subsequent analysis and augmentation.

Based on the selected attributes and user feedback, we applied domain-specific signal transformations and feature engineering to expand the dataset. This resulted in an augmented force profile set that preserved realism while increasing variability. All original and augmented profiles were then replayed on the simulator, and participants rated them using seven antonymous adjective

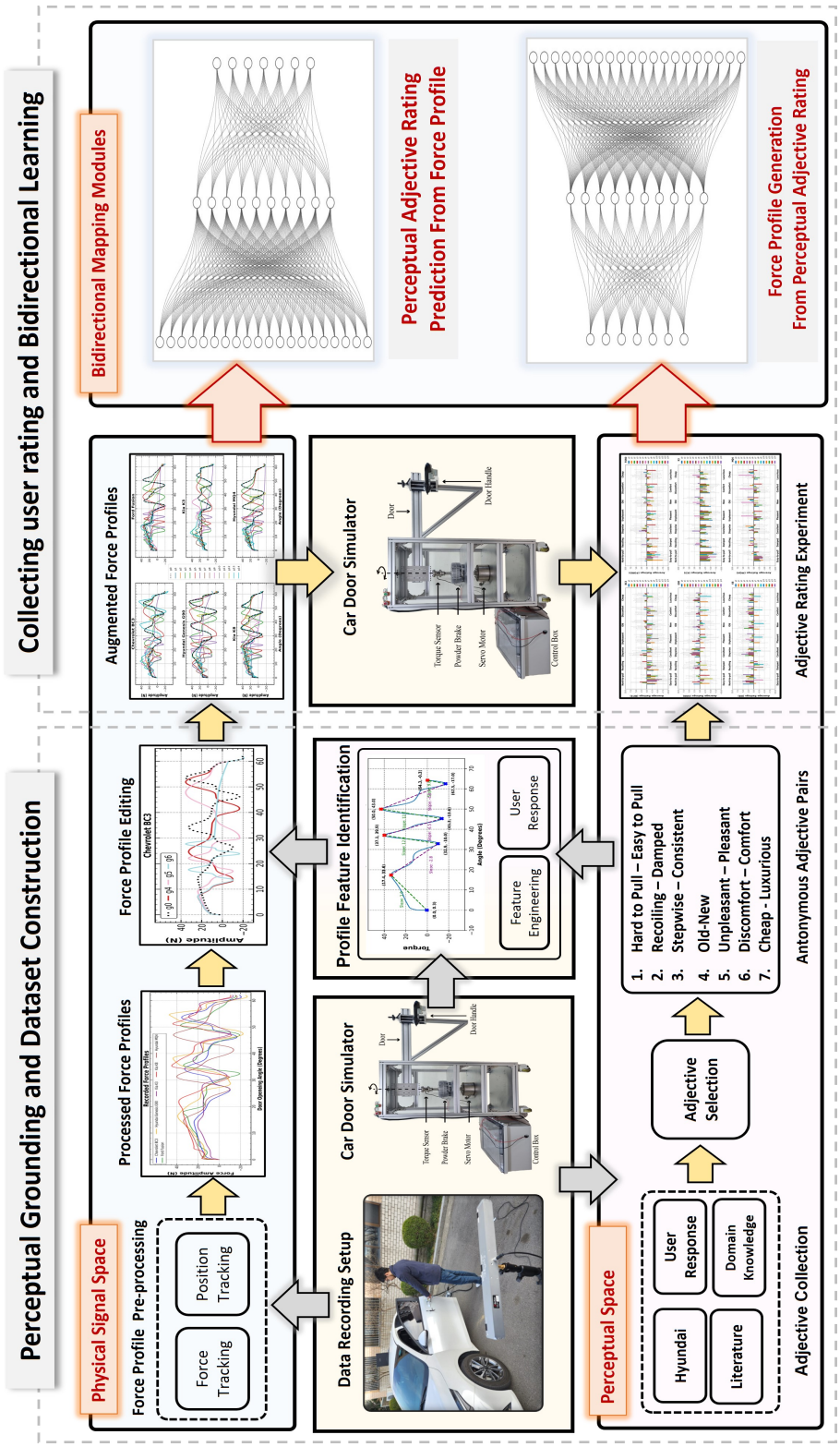


**Figure 5.1:** Overview of the proposed framework. Force profiles are first recorded from real car doors and processed into a normalized signal space. In parallel, perceptual attributes are identified through literature, expert input, and user studies with real cars and simulator playback. The data are augmented using signal-level transformations guided by user feedback and feature analysis. All profiles are replayed on a car door simulator, and users rate them using antonymous adjective pairs. These paired datasets are used to train bidirectional models for predicting perceptual ratings from signals and generating signals from desired ratings.

pairs. This process ensured consistency in haptic experience and minimized visual or contextual bias.

### 5.2.3 Bidirectional Modeling Approach

Once the Physical Signal Space and Perceptual Attribute Space were constructed and paired, we trained two models:

- A **Signal-to-Rating Model**, based on a CNN-LSTM architecture with residual connections, which predicts user ratings from a force profile. This model captures both local signal features and temporal dynamics to estimate perceptual responses.
- A **Rating-to-Signal Model**, which performs the inverse task of generating a physically plausible force profile from a given set of perceptual ratings. This model leverages the decoder portion of the pretrained Signal-to-Rating Model and uses it in conjunction with an additional encoder to map from the perceptual domain to the latent signal representation.

Together, these models form a closed-loop system that supports both evaluation and synthesis. Designers can either assess how a given signal is likely to be perceived or author a new signal that matches a desired perceptual intent.

## 5.3 Acquisition of Force Profiles and Perceptual Attribute Ratings

### 5.3.1 Force Profile of Opening a Car Door

In the psychophysical experiments, users opened a car door and provided perceptual ratings. The perceptual characteristics exhibited by an opening car door are highly dependent on the physical aspects of the door. Therefore, a physical signal that can describe the act of opening a door should be considered significant. The force profile can be considered an important physical aspect of opening a door. It refers to the amount of force required to open (and close) the door at different points in its range of motion.

It takes into account several factors that contribute to the perceptual characteristics of a car door. It can be considered as the combined effect of the weight of the door, its aerodynamics, and

the shape of the hinge that keeps it attached to the main frame. Therefore, it was decided to use the force profile for predicting the perceptual characteristics of opening a door. In the current study, force profiles of the cars provided in Section 5.3.1.2 were recorded. These were further processed to create uniform representations and then replayed using a programmable car door simulator to ensure consistent user interaction.

#### 5.3.1.1 Data Collection Setup

To record the force profile of the car door, we used an ATI force sensor and an Optitrack Trio120 optical sensor. The ATI force sensor was attached to the door handle, and Optitrack markers were placed just beside the handle so that they were visible to the cameras at all times. A one-time position tracking of the door hinge was carried out for every car. This was done to establish a reference point for measuring the opening angle. A user opened the door with their left hand. The users were instructed to make a conscious effort to maintain a constant velocity and avoid jerks. The force sensor recorded the force required to open the door at different points in its range of motion. The Optitrack Trio 120 was used to track the movement of the door and the markers to provide a visual representation of the door's range of motion. The complete setup is illustrated in Figure 5.2. The data from both sensors were synchronized based on timestamps. The force sensor recorded data at 1 kHz while Optitrack provided position data at an update rate of 80 Hz. The position data were upsampled to match the force sensor update rate.

These recordings served as the raw input for signal normalization and also as reference inputs for the simulator. All profiles, including augmented versions described in Section 5.3.4.1, were later rendered using the physical simulator for perceptual evaluation.

#### 5.3.1.2 1D Force Profiles

The data collected from different cars was inconsistent because it was collected by human users. The maximum opening angles of the cars were also variable. To make the data more comparable and accurate, it was important to normalize it and make it uniform across all cars.

The maximum opening angle for all the cars was capped at 62°, as most cars had a maximum opening angle below this limit. For cars with smaller maximum angles, data were zero-padded

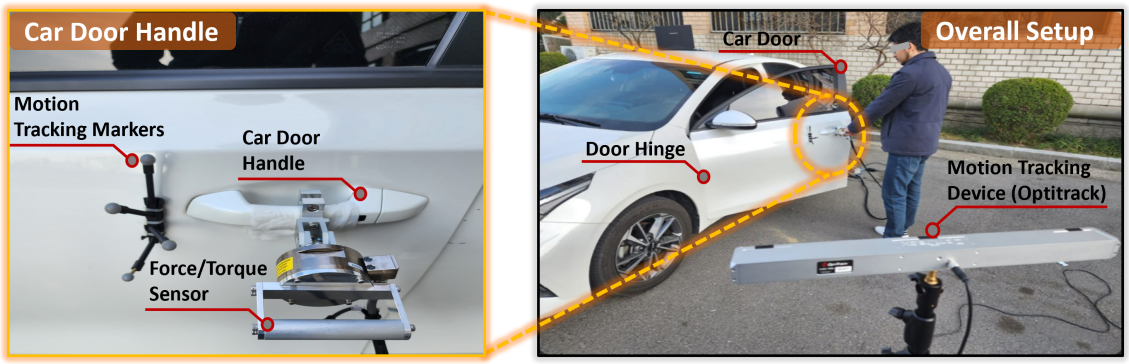


**Figure 5.2:** Experimental setup for recording car door force profiles and angular motion. A force sensor was attached to the door handle, and OptiTrack markers were placed to track the door's position during opening.

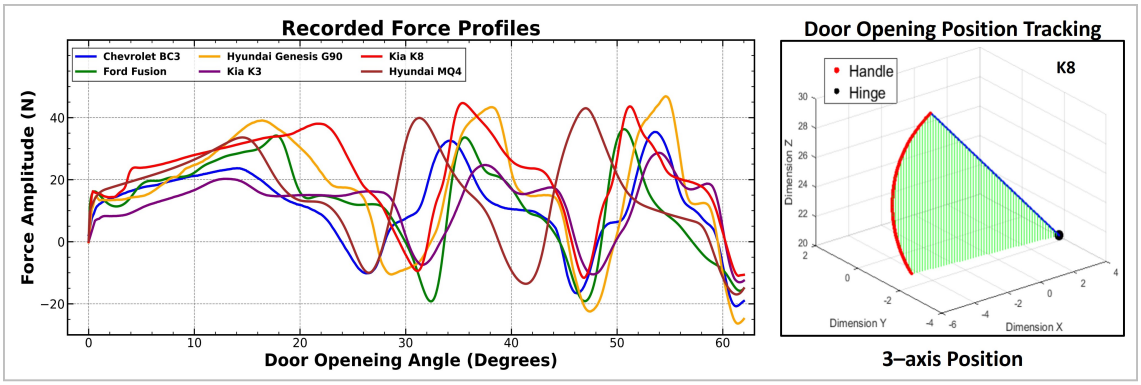


**Figure 5.3:** Angle-normalized force profiles of the six cars used in this study (Top). The position tracking of the door opening is provided for BC3, G90 and K8 for reference (Bottom).

at the end. Since data were collected by human users, the opening velocity was variable. This was normalized by combining the position tracking data and force data. The force data were divided into subsets corresponding to a range of 1° of the angle. The subset of force data for each degree was then downsampled and truncated to 10 data points. This was done to make the data uniform across all cars and smooth out outliers. This resampling process resulted in 621 data points per profile, accounting for the inclusion of the starting position at 0°, and provided a temporally uniform representation across all car profiles.

Force profiles of all six car doors and position tracking of Kia K8 car door is provided in Figure 5.3.

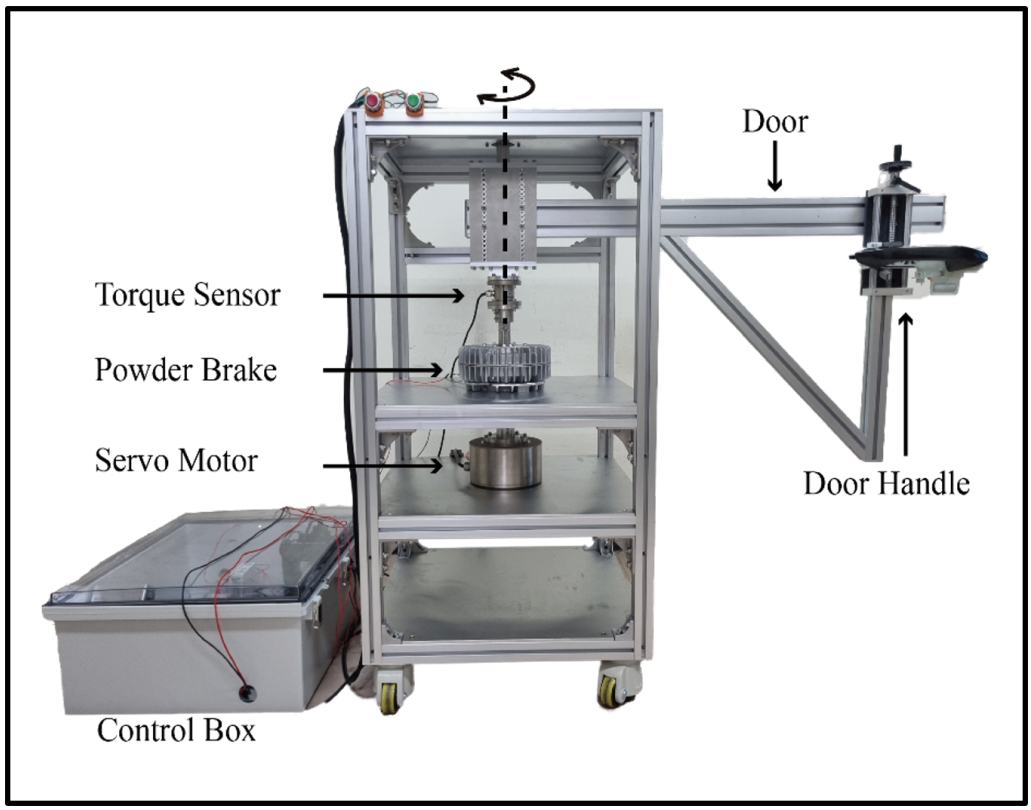


**Figure 5.4:** Car door simulator used for physical interaction. The system includes a direct-drive motor, magnetic powder brake, and angle encoder for high-fidelity torque playback.

### 5.3.2 Car Door Simulator

This study employed the hybrid car door simulator introduced by Ma et al. [108], which was specifically developed to replicate realistic automotive door interactions in a controlled experimental environment. The device consists of a single-degree-of-freedom (1-DoF) rotational arm actuated by a hybrid configuration of a direct-drive motor and a magnetic powder brake. The motor is responsible for replicating dynamic inertial and gravitational forces, while the brake contributes resistive torques corresponding to frictional or damping behavior.

The system is governed by a cascade control architecture with an inner velocity loop and an outer torque loop. A PID controller regulates torque feedback to ensure accurate tracking of target force profiles. The magnetic powder brake operates under open-loop current control, selectively engaging based on predefined torque states. An external rotary encoder continuously monitors angular position, enabling fine-grained playback of recorded or synthesized torque signals.

In the original study [108], the authors validated the simulator's fidelity using recorded data from actual car doors, achieving a high correlation ($r = 0.96$) and low RMSE (¡ 0.25 Nm) when replicating physical torque profiles. These results confirm the simulator's capability for precise reproduction of real-world mechanical characteristics, making it suitable for both psychophysical testing and force signal authoring.

Figure 5.4 illustrates the physical configuration of the device, showing the actuation components and participant interaction interface used in this study.

### 5.3.3 Perceptual Attribute Construction

This section outlines the structured process for developing a perceptual attribute space corresponding to the act of opening a car door. Three sequential user experiments were conducted: (1) collecting a perceptual adjective lexicon, (2) filtering for relevant attributes, and (3) rating physical signals along antonymous perceptual scales. Between Experiments 1 and 2, a targeted force profile augmentation phase was performed to expand the physical stimulus space based on user insight. All interactions in the second and third experiments used a calibrated car door simulator to ensure repeatability.

#### 5.3.3.1 Participants and Dataset

A total of 20 participants took part in the first and second experiments, and 26 in the third. Around 75% of the participants in all experiments were common, the remaining were replaced due to non-availability. The majority of the participants identified as males, while 10 out of the combined 66 across all experiments identified as females. Their average age was 27.5 years (range: 21 - 34). None of the participants reported any disabilities or any other factors that could prevent them from successfully participating in the experiments. All participants were compensated with $15 USD per experiment.

Six commercially available cars were used: BC3, FORD, G90, K3, K8, and MQ4. These cars were chosen to capture variation in door-opening dynamics and to span a spectrum from utility-class to luxury-class vehicles. Force profiles were collected using high-resolution sensors and used throughout this section. Augmented profiles were derived from these originals and replayed on

**Table 5.1:** The lexicon of adjectives built from four sources, i.e., Hyundai research (green), Experiment (black), literature (red), and domain expert (blue). The overall list was formed as a result of experiments 1 and 2.

| | | | |
|---|---|---|---|
| 1 Agitating | 18 Easy to operate | 35 Harmonic | 52 Cheerful, rhythmical |
| 2 Archaic | 19 Effortless | 36 Heavy | 53 Rigid |
| 3 Balanced | 20 Empty | 37 High | 54 Rough |
| 4 Calm | 21 Erratic | 38 Jarring | 55 Shaking |
| 5 Calming | 22 Exciting | 39 Jerky | 56 Shallow |
| 6 Cheap | 23 Expensive | 40 Joyful | 57 Smooth |
| 7 Classy | 24 Stepwise | 41 Light | 58 Soft |
| 8 Clinging | 25 Fluctuating | 42 Like new | 59 Sophisticated |
| 9 Comfortable | 26 Fluid | 43 Loud | 60 Stiff |
| 10 Consistent | 27 Forceful | 44 Luxurious | 61 Stressing |
| 11 Constant | 28 Free | 45 Natural | 62 Stuck |
| 12 Cool | 29 Frictional | 46 Not fit | 63 Tightly fit |
| 13 Damped | 30 Frictionless | 47 Old | 64 Uncomfortable |
| 14 Discordant | 31 Futuristic | 48 Pleasant | 65 Unpleasant |
| 15 Disturbing | 32 Gloomy | 49 Quiet | 66 Unstable |
| 16 Easy | 33 Hard | 50 Recoiling | 67 Vibrating |
| 17 Easy to open | 34 Hard to pull | 51 Relaxing | 68 Vintage |

the car door simulator for consistency across participants.

#### 5.3.3.2 Experiment 1: Adjective Lexicon Development

In the first experiment, participants opened the driver-side door of each real car and wrote down adjectives that best described their perceptual experience. They were instructed to focus on motion feel, resistance, smoothness, and overall tactile impression. No time limit was imposed, and the task was repeated for all six cars.

Additional adjectives were gathered from three sources: literature on haptic texture perception, domain expertise of the authors, and evaluation documents provided by Hyundai. This resulted in a broad lexicon reflecting affective, mechanical, and sensory attributes5.1.

#### 5.3.3.3 Experiment 2: Attribute Selection

In this phase, participants interacted with a randomized mix of real profiles played on the simulator. After each trial, they reviewed the 68-adjective list and marked those they felt described

the experience. Responses were binary (selected or not selected), and selection percentages were computed across users. Adjectives chosen by at least 20% of participants for any profile were retained for the final list.

#### 5.3.3.4 Results of Experiment 1 and 2

In the lexicon of adjectives, four different sources contributed adjectives. Among these sources, the user experiment provided a total of 33 unique adjectives. Hyundai uses adjectives for measuring the physical performance of a car door, eight of these were usable for our purpose. Thirteen adjectives were collected from prior literature [117, 118]. After analyzing the above three sources, the authors included 14 more adjectives based on their experience and knowledge of working in this domain. They felt these could be useful additions to the lexicon of adjectives. Combining all these sources, the lexicon contained a total of 68 adjectives, which are provided in Table 5.1.

The second experiment filtered out the most relevant adjectives for describing the perception of opening a door. Each adjective was scored by the users, and these scores were averaged for all cars and users. Figure 5.5 shows the relevance of each adjective. It was empirically decided to choose the adjectives that were selected by at least 20% of the users. A total of 25 out of the 68 adjectives were selected based on this criterion. These were further used in experiment 3.

### 5.3.4 Perceptual Adjective Ratings

#### 5.3.4.1 Dataset: Force Profile Augmentation

To enable a richer exploration of force-to-perception relationships, a set of augmented force profiles was created by integrating participant feedback with objective signal characteristics. Before implementing this augmentation, participants were interviewed about their general experiences with car doors, independent of the study context. These discussions consistently highlighted several perceptually salient features, such as the amplitude of the initial force peak, the timing and spacing between peaks, the presence and shape of plateaus, and changes in slope throughout the motion. Participants frequently associated these aspects with impressions of heaviness, resistance, and mechanical refinement.

To quantify these impressions, the original force profiles were analyzed using feature engineer-

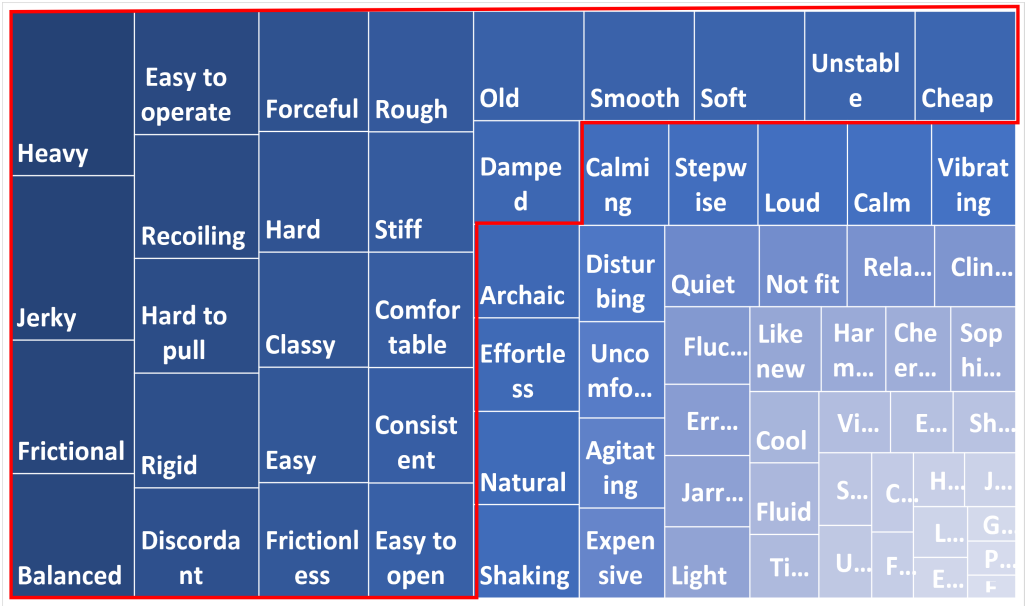


**Figure 5.5:** Relevance of all adjectives shown in percentage. The sizes of the boxes are sorted according to relevance percentage and the red border outlines the adjectives that were considered as relevant by at least 20% of the users.

ing techniques. Descriptors such as peak magnitude, angular intervals, and slope transitions were extracted and visualized (see Figure 5.6). This analysis helped bridge the gap between subjective perception and measurable signal traits, providing a foundation for structured augmentation.

Based on the combined insights from user interviews and feature analysis, a set of fourteen augmented variants was generated for each of the original profiles. This augmentation process was introduced after the initial experiments and resulted in a total of 90 unique profiles across six car models. Each profile was treated independently in the subsequent perceptual studies.

Five augmentation strategies were applied:

- **Amplitude Adjustment (g1–g3):** The first peak amplitude was modified to fixed values of 20 N, 30 N, and 40 N, simulating varying initial resistance.
- **Peak Position Adjustment (g4–g6):** The angular positions of the first three cycles were shifted earlier in the profile. Specifically, G4 compressed the first cycle to complete before 15°, G5 compressed the second to complete before 30°, and G6 did the same for the third.
- **Random Perturbations (g7–g9):** These modifications introduced localized changes such

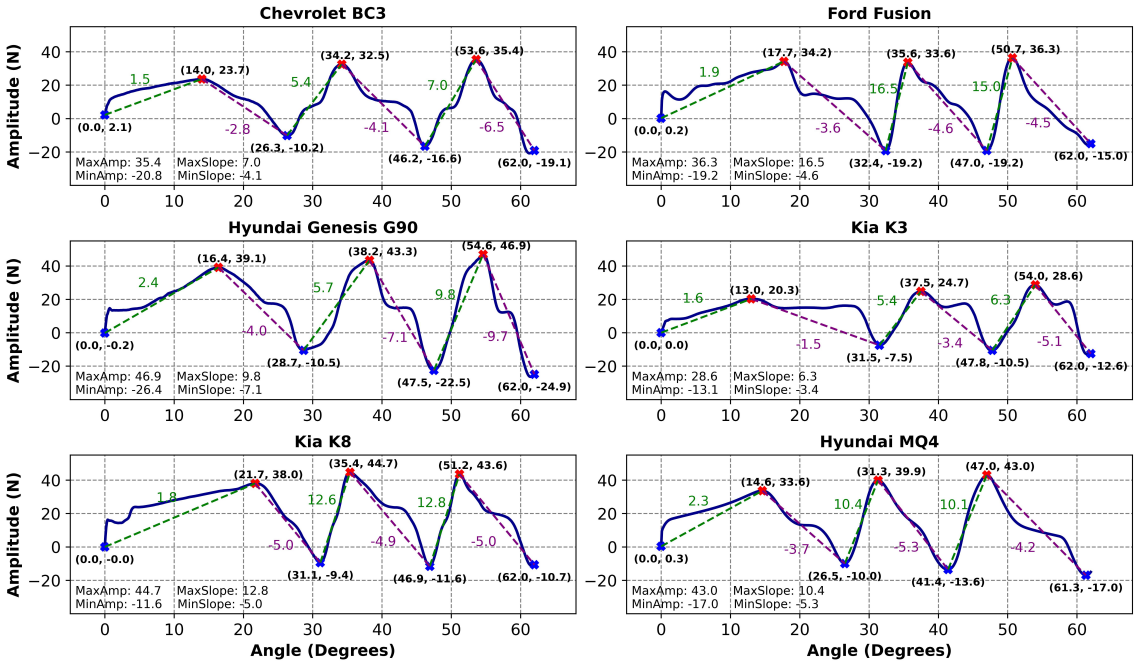


**Figure 5.6:** Processed force profiles of all six car models. Each plot highlights key engineered features used for augmentation, including peak magnitudes, inter-peak intervals, and slope transitions.

as moving the first peak to 10°, flattening sections up to 20°, inserting small cycles before the first peak, or between the first and second major transitions.

- **Plateau Modifications (g10–g12):** These versions altered plateau segments after the second peak, ranging from no change (g10), to stretched (g11), to both stretched and heightened (g12) plateaus.

- **Pre-Peak Bumps (g13–g14):** Small oscillatory features were added before the main peak using g2 and g3 as base profiles, mimicking anticipatory tactile events.

All augmented profiles were resampled to 621 time steps for temporal consistency and validated through replay using the car door simulator. These signals were subsequently integrated into Experiment 3 and treated equivalently to real measurements. This augmentation strategy preserved perceptual plausibility and expanded the dataset. It enabled a more controlled investigation of force-to-perception mappings while reducing reliance on additional real-world recordings. The complete set of augmented profiles is shown in Figure 5.7.

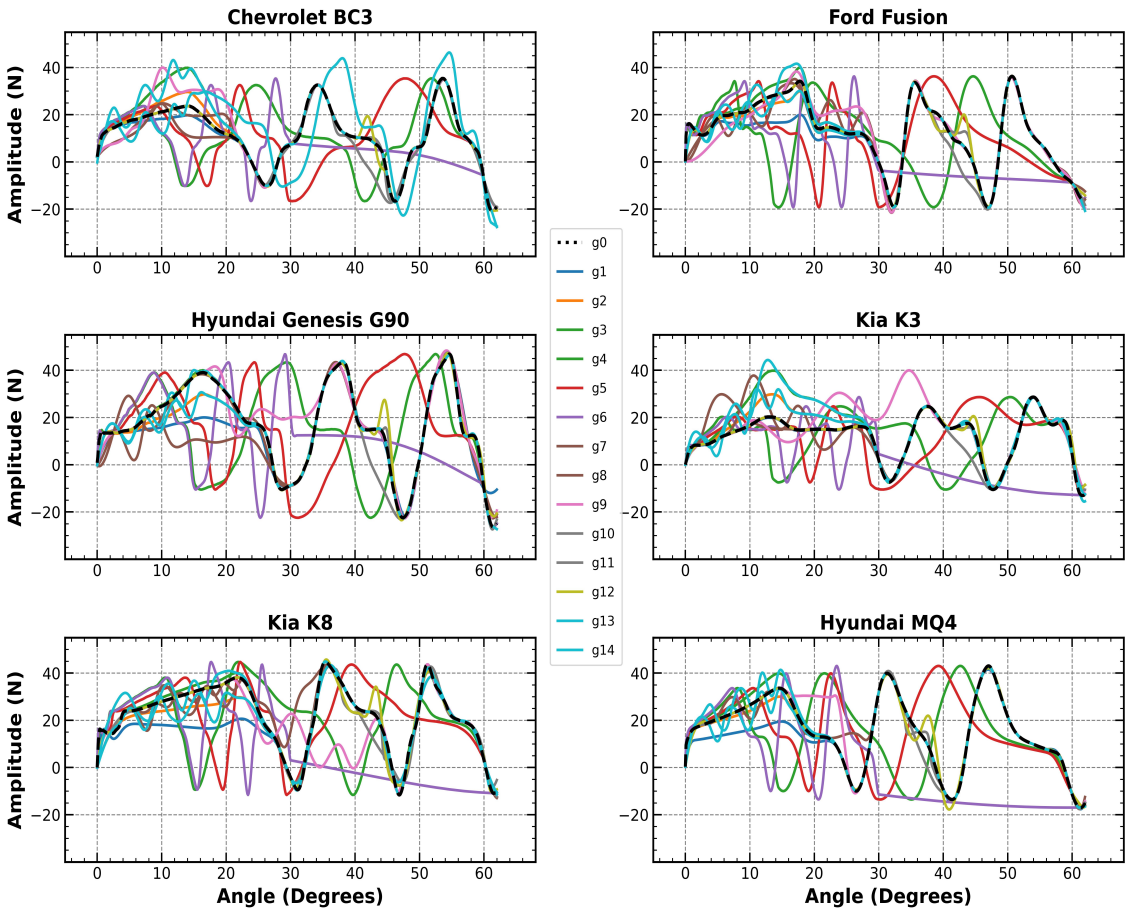


**Figure 5.7:** Visualization of all 90 force profiles used in the study, including 15 profiles per car. For each car, **g0** denotes the original recorded profile, while **g1** to **g14** represent the augmented variants. The augmented profiles span five groups based on manipulated characteristics: peak amplitude (g1–g3), peak position (g4–g6), random perturbations (g7–g9), plateau modifications (g10–g12), and pre-peak bumps (g13–g14).

#### 5.3.4.2 Experiment 3: Adjective Ratings

**Participants:**

A total of 40 participants (32 male, 8 female; mean age = 43.6 years) took part in this experiment. Among them, 40% had also participated in Experiments 1 and 2. All participants were asked to evaluate the perceived qualities of car door openings using a set of bidirectional adjective pairs. No participant reported any physical or cognitive conditions that could interfere with task performance or perceptual judgment.

**Stimuli:** Original force profiles from six car models (i.e., BC3, FORD, G90, K3, K8, and MQ4)

were used in this experiment. Each original profile was algorithmically modified to generate 14 augmented versions, resulting in 15 profiles per car and 90 total. To mitigate fatigue, profiles were split into two sets. Group 1 participants evaluated profiles from BC3, FORD, and G90, while Group 2 participants evaluated those from K3, K8, and MQ4. This division also reduced perceptual overlap, as similar profiles such as BC3 and K3 or G90 and K8 were intentionally placed in different groups.

**Adjective Pairs:** From the 25 adjectives refined through earlier filtering and participant input, seven representative pairs were chosen. These pairs captured both physical interaction qualities (e.g., Easy-to-pull vs. Hard-to-pull, Damped vs. Recoiling, Consistent vs. Stepwise) and affective impressions (e.g., Pleasant vs. Unpleasant, New vs. Old, Comfort vs. Discomfort, Luxurious vs. Cheap). Reducing the attribute set in this way minimized cognitive burden while preserving the richness of perceptual representation.

**Table 5.2:** Seven adjective pairs used in the adjective rating experiment.

| # | Attribute | Antonym |
|---|---|---|
| 1 | Easy-to-pull | Hard-to-pull |
| 2 | Damped | Recoiling |
| 3 | Consistent | Stepwise |
| 4 | Pleasant | Unpleasant |
| 5 | New | Old |
| 6 | Comfort | Discomfort |
| 7 | Luxurious | Cheap |

**Graphical User Interface for Adjective Ratings.** A custom graphical user interface (GUI) was developed to collect participant responses (see Figure 5.8). The interface displayed each adjective pair on a continuous horizontal scale ranging from –10 to +10, with a resolution of 1. Each extreme corresponded to a strong perception of the associated adjective, while zero indicated a neutral impression. Participants submitted their ratings using a mouse. The GUI also supported forward and backward navigation, enabling participants to revisit and revise their responses if necessary.

In addition to collecting perceptual ratings, the GUI facilitated experimental control and session tracking. It allowed the experimenter to enter participant information, including age, gender, and driving experience. The interface also supported assignment to experimental groups and selection of pre-defined Latin sequences for randomized profile presentation. These features ensured

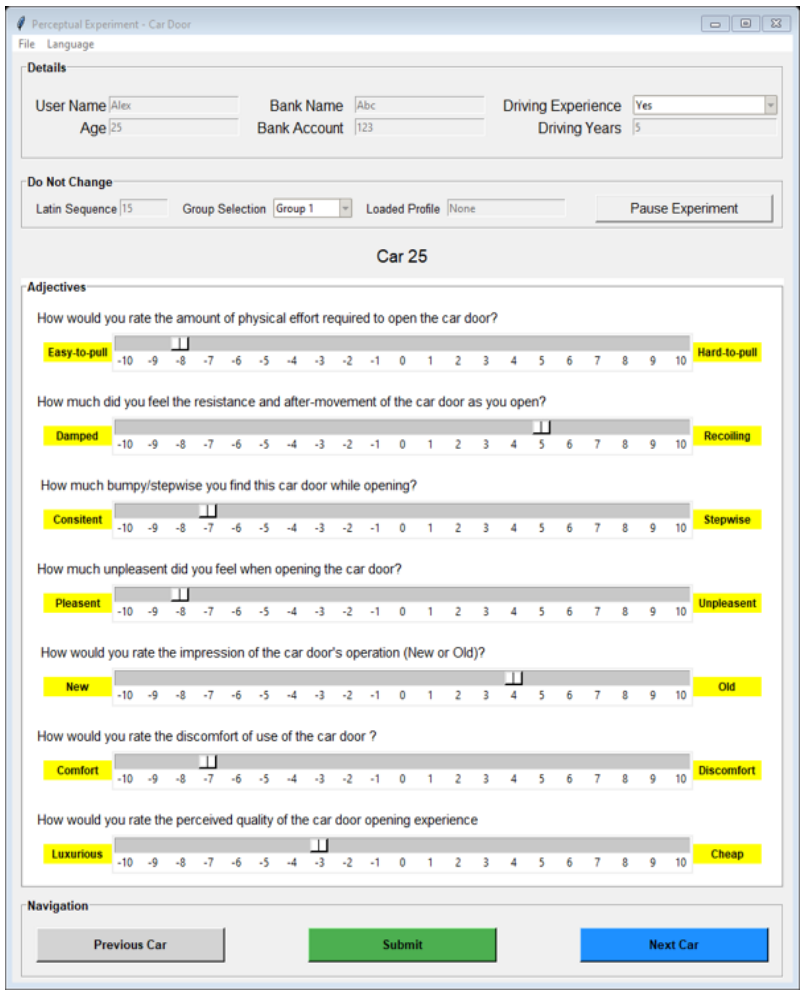


**Figure 5.8:** Graphical user interface (GUI) used for the adjective rating experiment.

consistent protocol execution and streamlined administration across all participants.

**Experimental Setup.** The experiment was conducted using a car door simulator designed to replicate natural door-opening dynamics. Participants received both written and verbal instructions prior to the session. During each trial, they wore headphones and were blindfolded to minimize auditory and visual distractions. Each participant opened the car door fully using their dominant hand in a natural motion. Force profiles were presented in a randomized order, as defined by a pre-generated Latin sequence loaded through the GUI. Participants entered their responses after each trial via the interface, which remained active throughout the session. Breaks were allowed at the participants' discretion, and each session lasted approximately 90 minutes.

#### 5.3.4.3 Results and Analysis

The data collected in Experiment 3 consisted of user ratings across seven adjective pairs for each of the 90 force profiles. All responses were normalized to a common scale ranging from –10 to +10. For each profile, user ratings were averaged across participants to compute a representative perceptual score. Mean values and standard deviations were then calculated over all 40 participants for each adjective and profile group. The resulting distribution is visualized in Figure 5.9, where error bars represent inter-participant variability.

To investigate how users perceived the different augmented profiles, ratings were further analyzed by grouping the profiles according to the applied augmentation strategy. The seven adjective pairs were categorized into physical and emotional perceptual attributes. Figure 5.10 presents the average ratings for each augmentation group and attribute category. Each group of augmented profiles (e.g., g1 to g3 for amplitude scaling) exhibited interpretable perceptual trends across both physical and emotional dimensions. Increasing the amplitude of the initial peak from 20 N to 40 N (g1–g3) consistently elevated ratings for "Hard-to-pull," "Recoiling," and "Stepwise," confirming the influence of early peak strength on perceived effort and control. The compression group (g4–g6), which altered the angular timing of peak occurrences, showed the most pronounced directional changes. In particular, g6 produced a distinct shift in "Recoiling" ratings. Because all major peaks occurred before 30 degrees, the force increased rapidly toward the end of the opening motion, resulting in a strong pull that was perceived as an abrupt recoil.

Interestingly, this sharp terminal sensation also influenced emotional impressions: participants rated g6 profiles lower on attributes such as "Luxurious" and "Pleasant," often associating them with lower-end or less refined experiences. This suggests that uncontrolled or sudden force dynamics may degrade perceived quality. Modifications in plateau structure (g10–g12) produced moderate yet consistent rating changes, indicating that extended or heightened plateaus mildly influence perceptions of smoothness and control. The final group (g13–g14), which introduced small pre-peak bumps, resulted in measurable shifts in both physical and emotional attributes, emphasizing user sensitivity to early force transitions. As expected, randomly perturbed profiles (g7–g9) lacked directional trends, reinforcing the benefit of structured feature modifications for producing predictable perceptual outcomes.

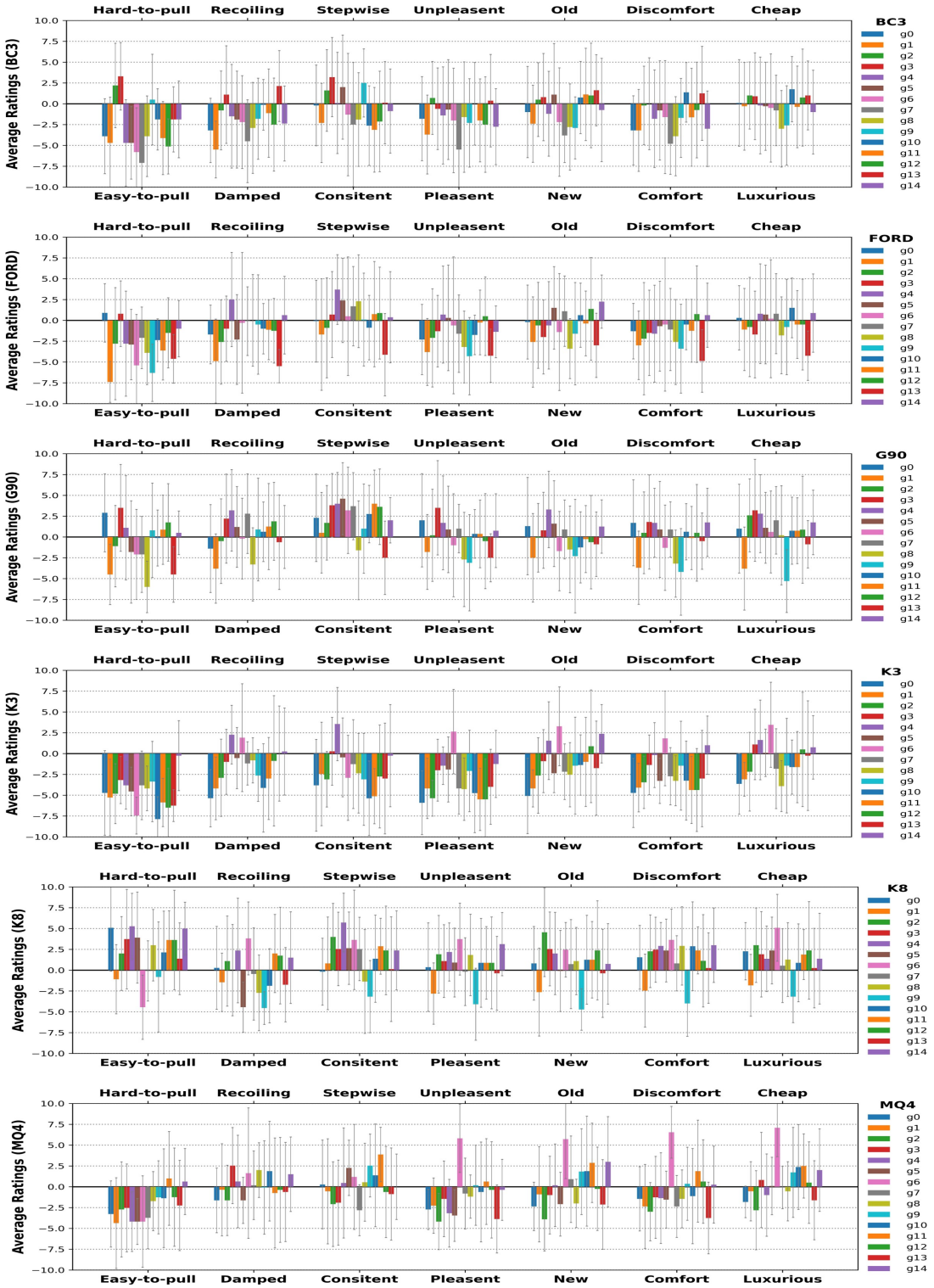


**Figure 5.9:** Average adjective ratings across original and augmented profiles. Bars represent mean scores and error bars denote standard deviations across participants.

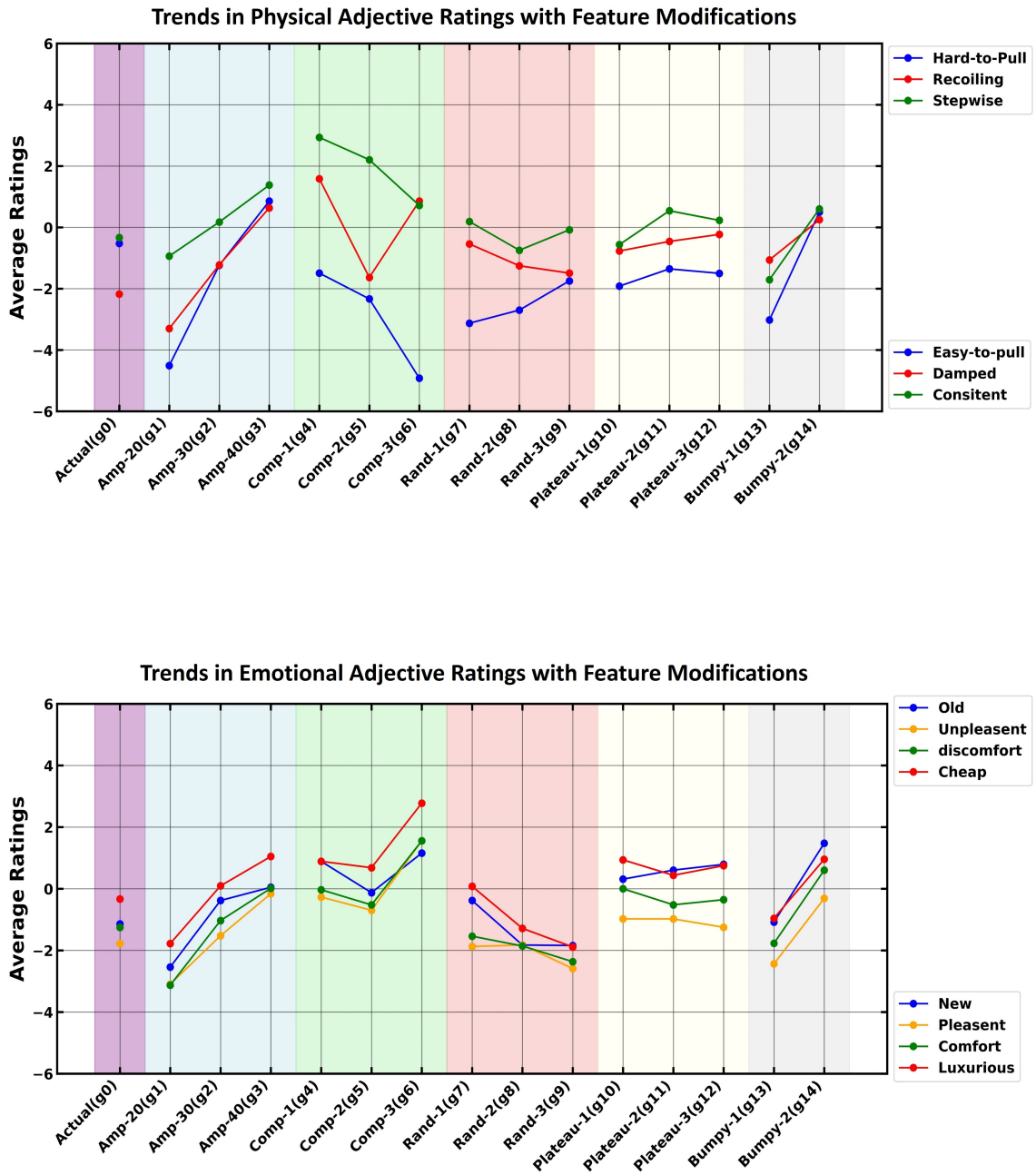


**Figure 5.10:** Perceptual trends across augmentation groups. (Top) Physical attribute ratings show consistent within-group variation across amplitude, compression, plateau, and bump modifications. (Bottom) Emotional attribute ratings reflect corresponding perceptual shifts, while randomly perturbed profiles (g7–g9) show no directional trend.

## 5.4 Force-to-Perception Modeling

Traditional time-series models, such as Auto-Regressive (AR), Moving Average (MA), and their variants, often fail to capture the non-linear and hierarchical dependencies present in real-world perceptual signals. While recurrent architectures such as LSTMs can overcome vanishing gradient issues and learn long-term dependencies, they may be limited in extracting localized spatial patterns inherent in mechanical force profiles. In contrast, convolutional neural networks (CNNs) are well suited for identifying localized patterns and learning shift-invariant features through weight sharing. Recent research has demonstrated the utility of CNNs in various haptic learning tasks, including tactile attribute estimation and perceptual similarity modeling [19, 119].

In this study, a CNN-based encoder–decoder architecture with residual connections is proposed to predict perceptual ratings from force profiles collected during car door opening. The residual structure facilitates the training of deeper networks by mitigating vanishing gradient problems, while the encoder–decoder configuration enables effective feature compression and reconstruction, aligning with the sequential nature of the task. The model is optimized for regression using the Huber loss, and the output corresponds to a seven-dimensional vector representing user ratings on bipolar adjective scales.

### 5.4.1 Network Architecture

The proposed model adopts an encoder–decoder architecture designed to map one-dimensional force profiles to user-rated perceptual attributes. This formulation enables learning hierarchical representations from mechanical signals using a combination of convolutional and residual operations. The encoder compresses input signals into a latent feature space, while the decoder reconstructs perceptual ratings from this compact representation. An overview of the architecture is illustrated in Figure 5.11.

#### 5.4.1.1 Encoder Module

The encoder network is designed to transform a univariate force profile of length 621 into a compact latent representation that preserves perceptually relevant dynamics. The input sequence $\mathbf{x} \in \mathbb{R}^{621\times 1}$ is first passed through a one-dimensional convolutional layer with 64 filters of kernel

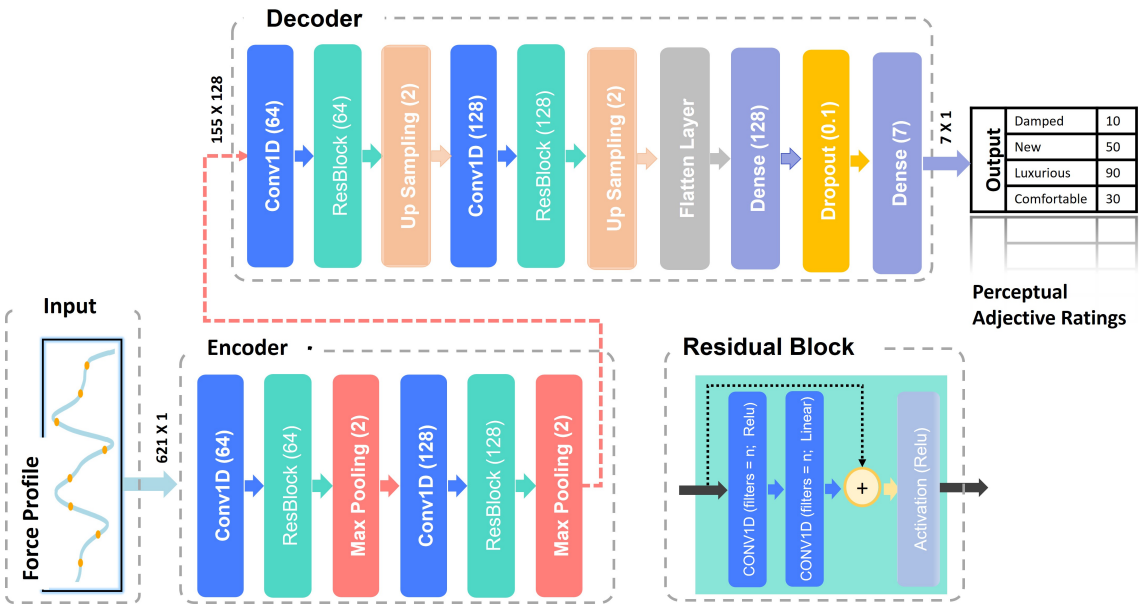


**Figure 5.11:** Architecture of the proposed 1D CNN encoder–decoder model with residual connections. The network takes a force profile as input and predicts the corresponding perceptual attribute ratings.

size 3 and ReLU activation. This operation captures localized temporal features and is formally defined as:

$$\mathbf{y}_t = \sigma\left(\sum_{i=0}^{k-1} \mathbf{w}_i \cdot \mathbf{x}_{t+i} + b\right) \tag{5.1}$$

where $\mathbf{w}_i$ are the convolution weights, $b$ is the bias term, $k$ is the kernel size, and $\sigma$ is the ReLU function.

To facilitate stable gradient flow and improve feature propagation, a residual block is applied. It consists of two convolutional layers with linear and ReLU activations, followed by a skip connection:

$$\mathbf{y} = \mathrm{ReLU}\left(\mathbf{x} + \mathrm{Conv}_2\left(\mathrm{ReLU}(\mathrm{Conv}_1(\mathbf{x}))\right)\right) \tag{5.2}$$

This is followed by a max pooling layer of size 2 to downsample the temporal resolution. The process is repeated using 128 filters in the second convolutional stage. After two downsampling stages, the temporal dimension is reduced from 621 to 155, and the feature dimensionality is increased to 128, resulting in a latent representation $\mathbf{z} \in \mathbb{R}^{155\times128}$.

#### 5.4.1.2 Decoder Module

The decoder takes the latent tensor $\mathbf{z}$ and reconstructs the target perceptual ratings. A 1D convolutional layer with 128 filters is followed by a residual block, then an upsampling layer doubles the temporal resolution. This is repeated using 64 filters and a second residual block, restoring the time axis to 620. The output is flattened and passed through a dense layer with 128 units, followed by a dropout layer with rate 0.1 for regularization. Finally, a linear activation layer produces a 7-dimensional output:

$$\hat{\mathbf{y}} = f_\theta(\mathbf{x}) \in \mathbb{R}^7 \tag{5.3}$$

Here, $f_\theta$ denotes the full encoder–decoder model, and $\hat{\mathbf{y}}$ represents the predicted perceptual attribute vector.

The network uses ReLU activations throughout all intermediate layers to promote sparsity and non-linearity, while the final output layer employs a linear activation to support regression over continuous perceptual scales.

#### 5.4.1.3 Training Objective

The model is trained to minimize the Huber loss, which combines the advantages of Mean Squared Error (MSE) and Mean Absolute Error (MAE). It is robust to outliers while maintaining sensitivity to small deviations:

$$\mathcal{L}_\delta(\hat{y}, y) = \begin{cases} \frac{1}{2}(\hat{y} - y)^2 & \text{if } |\hat{y} - y| \leq \delta \\ \delta \cdot \left(|\hat{y} - y| - \frac{1}{2}\delta\right) & \text{otherwise} \end{cases} \tag{5.4}$$

A threshold $\delta = 2.0$ was selected based on validation performance.

#### 5.4.1.4 Model Training Procedure

The model was trained to predict perceptual ratings across seven attributes ($\mathbb{R}^{7\times1}$) using a force profile input represented as a time series of shape $(621, 1)$. The training objective was to learn a direct mapping from the force signal to the corresponding perceptual ratings.

Multiple loss functions were evaluated, including Mean Squared Error (MSE), Mean Absolute

Error (MAE), and Huber loss. Among these, the Huber loss with $\delta = 2.0$ consistently yielded the most stable convergence and lowest validation error. The model was trained using the Adam optimizer for 200 epochs with a batch size of 8. A learning rate scheduler was employed to reduce the learning rate upon validation loss plateauing, with a minimum learning rate set to $10^{-6}$.

The network was implemented and trained using the TensorFlow Keras library. The resulting residual CNN-based encoder–decoder architecture is modular, interpretable, and well suited for signal-to-rating prediction in haptic perception tasks. It also provides a flexible foundation for future extensions involving cross-modal integration or attention-based mechanisms.

## 5.5 Evaluation

### 5.5.1 Dataset Preparation

The dataset used in this study consisted of 90 force profiles (15 per car) derived from six car models: BC3, FORD, G90, K3, K8, and MQ4. Each force profile was sampled at 621 time steps and paired with corresponding user-rated perceptual attributes spanning seven bipolar adjective pairs.

To ensure consistent model input, all force profiles and perceptual ratings were min–max normalized to a [0, 100] scale. Scaling was performed using global minimum and maximum values across the entire dataset, and the scaling parameters were stored for inverse transformation.

To enhance the diversity and robustness of the dataset, five additional variants were synthesized for each original profile using controlled augmentation techniques, including noise injection, time shifting, local smoothing, amplitude scaling, and spike insertion. Each augmented force profile was paired with an adjusted perceptual rating to reflect the expected perceptual shifts based on the nature of augmentation. This resulted in a significantly expanded dataset suitable for robust model training and evaluation.

### 5.5.2 Cross-Validation Strategy

To rigorously assess model generalization and ensure unbiased performance evaluation, a 5-fold cross-validation strategy was adopted. Unlike prior studies that rely on leave-one-out cross-

validation (LOOCV), which may produce optimistic results on small datasets, the proposed protocol ensures more reliable estimation by partitioning the data into training and test splits with balanced content.

In each fold, the dataset was randomly divided such that:

- 80% of the profiles (real + augmented) were used for training and validation,
- 20% were held out as the test set.

Care was taken to maintain balanced representation across all car models and augmentation types (g0 to g14) in each fold. Each force profile was used exactly once for testing, and four times for training, ensuring fair coverage of the dataset. The predicted ratings from each test fold were stored for further error analysis.

### 5.5.3 Evaluation Metrics

Three standard metrics were employed to assess the prediction accuracy of the proposed model:

- **Mean Absolute Error (MAE)**: Captures the average absolute difference between predicted and true values.
- **Root Mean Square Error (RMSE)**: Penalizes larger errors more strongly, emphasizing variance in prediction.
- **Coefficient of Determination ($R^2$)**: Measures the proportion of variance in the ground truth explained by the model.

$$\text{MAE} = \frac{1}{Md} \sum_{i=1}^{M} \sum_{j=1}^{d} |y_{ij} - \hat{y}_{ij}| \tag{5.5}$$

$$\text{RMSE} = \sqrt{\frac{1}{Md} \sum_{i=1}^{M} \sum_{j=1}^{d} (y_{ij} - \hat{y}_{ij})^2} \tag{5.6}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{M} \|y_i - \hat{y}_i\|_2^2}{\sum_{i=1}^{M} \|y_i - \bar{y}\|_2^2} \tag{5.7}$$

Here, $M$ denotes the number of test profiles, $d = 7$ represents the dimensionality of perceptual attributes, and $\bar{y}$ is the mean of the ground truth labels.

### 5.5.4 Performance Aggregation and Analysis

To analyze model performance comprehensively, results were aggregated along the following dimensions:

- **Overall**: Averaged across all 90 profiles (real + augmented) to provide a general performance estimate.
- **Per Car Model**: MAE, RMSE, and $R^2$ were computed separately for each car by averaging over its 15 profiles.
- **Per Profile Type (g0–g14)**: Metrics were computed across all six cars for each augmentation type to assess the model's sensitivity and consistency across engineered variations.

During evaluation, each test prediction was logged with the corresponding car ID and profile variation (e.g., 'BC3_g4') to enable targeted error analysis and generation of trend visualizations (e.g., MAE per g-group). This allowed the model's capacity to generalize across both mechanical and perceptual diversity to be rigorously quantified.

All results were obtained by averaging test performance across the five folds. These aggregated outcomes form the basis of the numerical and perceptual analysis presented in the subsequent section.

#### 5.5.4.1 Model Performance

The trained CNN encoder-decoder model was evaluated using 5-fold cross-validation across both real and augmented car door profiles. The prediction performance was assessed in terms of Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and the coefficient of determination ($R^2$), as shown in Tables 5.3, 5.4, and 5.5, respectively. These tables provide a breakdown of prediction performance across the six car models (BC3, FORD, G90, K3, K8, MQ4) and all seven adjective pairs.

**Table 5.3:** Average $MAE$ scores for each car and its 14 augmented variants (6 cars $\times$ 15 profiles = 90) used in this study. The predicted and human-rated values are reported for seven adjective pairs.

| **Attribute** | **BC3** | **FORD** | **G90** | **K3** | **K8** | **MQ4** | **Avg. MAE %** |
|---|---|---|---|---|---|---|---|
| Easy-to-pull / Hard-to-*pull* | 5.0 | 4.65 | 2.39 | 3.69 | 3.37 | 2.12 | **3.54** |
| Damped / Recoiling | 5.61 | 2.48 | 3.31 | 1.85 | 1.28 | 2.68 | **2.87** |
| Consistent / Stepwise | 4.44 | 1.73 | 2.32 | 3.29 | 3.14 | 3.52 | **3.07** |
| Pleasant / Unpleasant | 3.94 | 1.63 | 1.96 | 2.88 | 1.74 | 2.21 | **2.39** |
| New / Old | 4.39 | 3.82 | 1.58 | 1.79 | 2.92 | 3.14 | **2.94** |
| Comfort / Discomfort | 3.37 | 2.01 | 2.03 | 1.6 | 2.11 | 1.97 | **2.18** |
| Luxurious / Cheap | 3.97 | 3.34 | 3.61 | 2.32 | 2.14 | 2.72 | **3.02** |
| **Average (Car-wise)** | **4.39** | **2.81** | **2.45** | **2.49** | **2.39** | **2.62** | **—** |

The Mean Absolute Error (MAE) was calculated for all the adjective pairs and all the cars to better understand the prediction results, as shown in Table 5.3. The MAE offers a more direct and intuitive summary of the prediction results. Table 5.3 shows the individual prediction accuracy for each car against each of the adjective pairs. The MAE $\%$ column on the right shows the averaged prediction error for each car, while the MAE $\%$ column at the bottom shows the averaged prediction error for each adjective pair. It can be seen that the average prediction accuracy for most of the cars and adjective pairs is below 10 $\%$ a JND used in similar studies.

As seen in Table 5.3, the average MAE across most cars and attributes remained below 3.5%, suggesting that the model could generalize perceptual judgments from force profiles with high accuracy. The highest individual MAE was observed for K3 (4.39%), which aligns with its outlier status in force dynamics. The lowest MAE was achieved on MQ4 (2.39%), indicating more stable prediction performance on cars with smoother or more typical force profiles.

Attribute-wise, the pair "Damped–Recoiling" achieved the lowest average MAE (2.87%), followed by "Pleasant–Unpleasant" (2.39%) and "Comfort–Discomfort" (2.18%). These results suggest that users' perception of damping and comfort-related cues can be reliably inferred from the shape and amplitude of the force profile. In contrast, the attribute "Luxurious–Cheap" yielded a higher average MAE (3.02%), reflecting the cognitive ambiguity and subjectivity associated with interpreting luxury from physical interaction alone.

The RMSE values shown in Table 5.4 follow a similar trend, reinforcing the consistency of the error margins. The best overall performance was again found for MQ4 (3.23%) and K8 (3.43%),

**Table 5.4:** Average $RMSE$ scores for each car and its 14 augmented variants (6 cars $\times$ 15 profiles = 90) used in this study. The predicted and human-rated values are reported for seven adjective pairs.

| **Attribute** | **BC3** | **FORD** | **G90** | **K3** | **K8** | **MQ4** | **Avg. RMSE %** |
|---|---|---|---|---|---|---|---|
| Easy-to-pull / Hard-to-*pull* | 7.07 | 7.21 | 4.14 | 4.98 | 4.01 | 2.99 | **5.07** |
| Damped / Recoiling | 8.39 | 3.03 | 4.73 | 2.57 | 1.87 | 4.28 | **4.14** |
| Consistent / Stepwise | 5.94 | 2.36 | 3.27 | 5.12 | 4.36 | 5.37 | **4.40** |
| Pleasant / Unpleasant | 5.32 | 2.09 | 2.59 | 3.99 | 2.46 | 3.53 | **3.33** |
| New / Old | 5.18 | 4.68 | 2.03 | 2.03 | 4.80 | 3.92 | **3.77** |
| Comfort / Discomfort | 4.45 | 2.61 | 2.38 | 2.02 | 2.39 | 2.81 | **2.78** |
| Luxurious / Cheap | 6.02 | 4.34 | 4.98 | 3.29 | 2.69 | 4.41 | **4.29** |
| **Average (Car-wise)** | **6.05** | **3.76** | **3.44** | **3.43** | **3.23** | **3.90** | **—** |

**Table 5.5:** Average $R^2$ scores for each car and its 14 augmented variants (6 cars $\times$ 15 profiles = 90) used in this study. The predicted and human-rated values are reported for seven adjective pairs.

| **Attribute** | **BC3** | **FORD** | **G90** | **K3** | **K8** | **MQ4** | **Avg. $R^2$** |
|---|---|---|---|---|---|---|---|
| Easy-to-pull / Hard-to-*pull* | 0.82 | 0.77 | 0.95 | 0.86 | 0.93 | 0.93 | **0.88** |
| Damped / Recoiling | 0.63 | 0.95 | 0.87 | 0.96 | 0.99 | 0.74 | **0.86** |
| Consistent / Stepwise | 0.80 | 0.96 | 0.95 | 0.86 | 0.89 | 0.82 | **0.88** |
| Pleasant / Unpleasant | 0.71 | 0.96 | 0.95 | 0.92 | 0.96 | 0.95 | **0.91** |
| New / Old | 0.80 | 0.83 | 0.96 | 0.98 | 0.87 | 0.96 | **0.90** |
| Comfort / Discomfort | 0.86 | 0.93 | 0.97 | 0.97 | 0.97 | 0.97 | **0.94** |
| Luxurious / Cheap | 0.50 | 0.79 | 0.89 | 0.92 | 0.95 | 0.93 | **0.83** |
| **Average (Car-wise)** | **0.73** | **0.88** | **0.93** | **0.92** | **0.94** | **0.90** | **—** |

whereas BC3 showed the highest average RMSE (6.05%). Among the adjective pairs, "Comfort–Discomfort" had the lowest RMSE (2.78%) and "Easy-to-pull" the highest (5.07%). Notably, prediction variance remained within perceptual limits, which is further confirmed by the $R^2$ values. Table 5.5 shows that the $R^2$ scores exceeded 0.90 for most attributes, highlighting strong predictive alignment with user ratings. The highest $R^2$ was observed for "Comfort–Discomfort" (0.94), while "Luxurious–Cheap" and "Damped–Recoiling" showed slightly lower agreement (0.83 and 0.86, respectively). The model achieved a global average $R^2$ of 0.90, confirming that a substantial proportion of the perceptual variance could be explained by the predicted force-driven features.

#### 5.5.4.2 Error Analysis

Figures 5.12 and 5.13 present a scatter analysis of the predicted versus actual ratings, highlighting the model's performance across car profiles and perceptual attributes, respectively. In both plots, the dashed diagonal represents the ideal prediction line (line of identity), while the colored bands indicate regions within $\pm 0.5$ and $\pm 1$ standard deviation of user ratings.

A majority of the data points fall within the $\pm 1$ standard deviation band, demonstrating that the predicted ratings closely align with participant responses. The global standard deviation of user ratings across the entire dataset was 16.4, which serves as a reference for the shaded bands. Points located above the identity line signify underestimation of user ratings, while those below indicate overestimation.

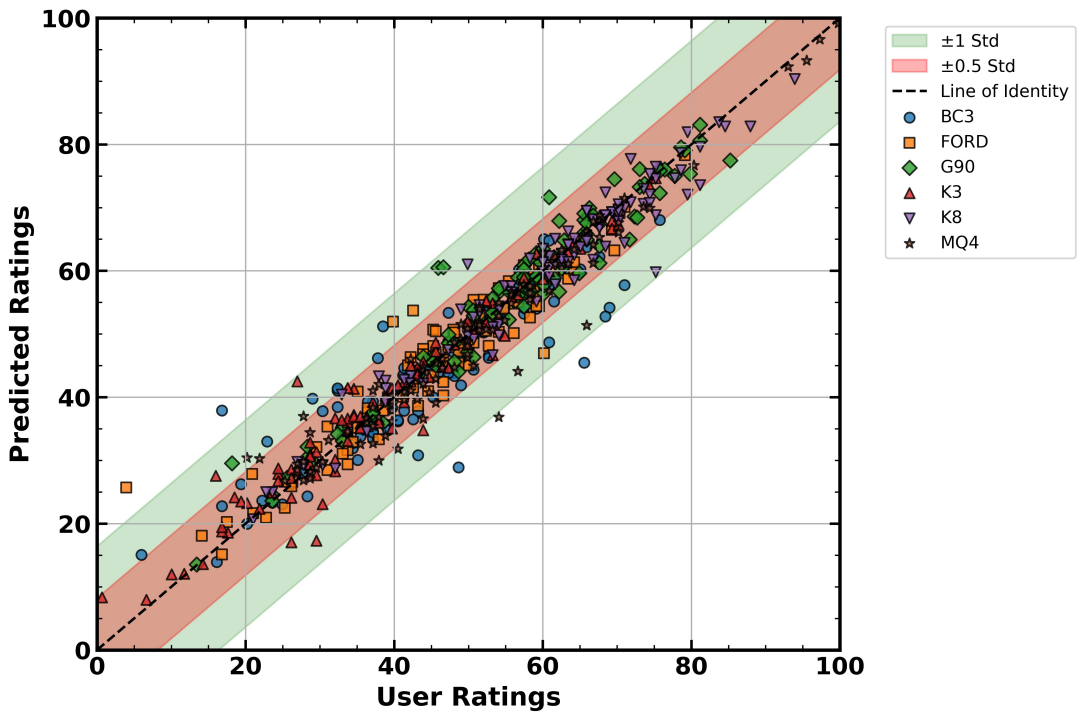


**Figure 5.12:** Prediction analysis based on user rating variability across different car profiles. The dashed line represents the ideal prediction (line of identity), while the red and green bands indicate the $\pm 0.5$ and $\pm 1$ standard deviation ranges of the user ratings, respectively.

When grouped by car model (Figure 5.12), prediction errors were generally consistent across vehicles, with slightly tighter clusters observed for MQ4 (STD = 15.98 actual, 15.33 predicted) and G90 (STD = 15.01 actual, 14.42 predicted). The lowest prediction deviation was observed

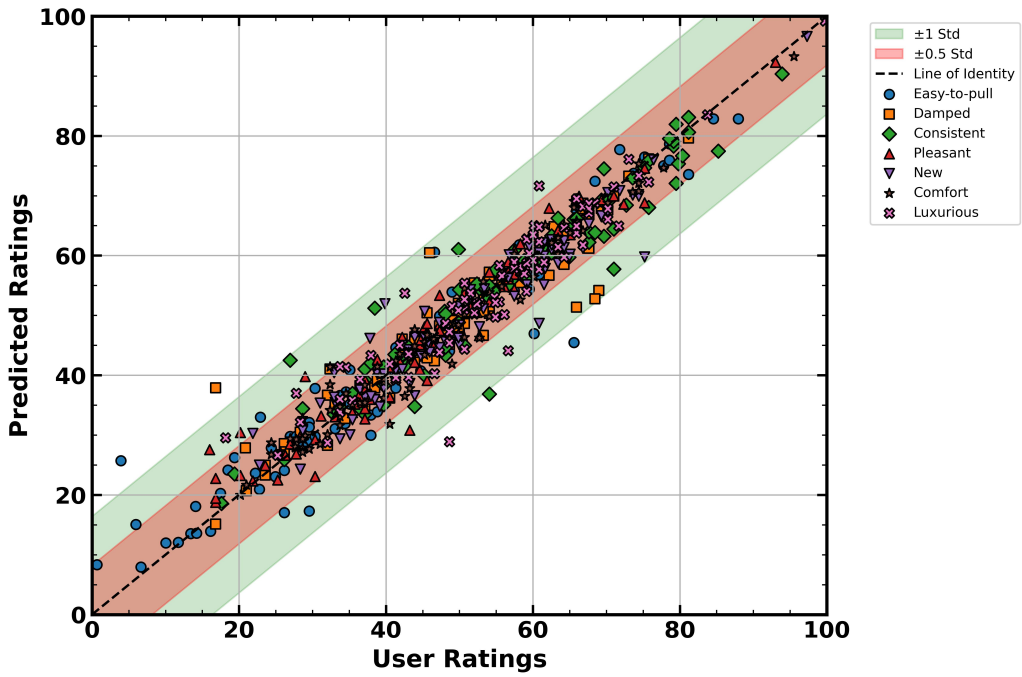


**Figure 5.13:** Prediction analysis based on user rating variability for each adjective pair. The dashed line represents the ideal prediction (line of identity), while the red and green bands denote the $\pm 0.5$ and $\pm 1$ standard deviation ranges of the user ratings, respectively.

for BC3, where the standard deviation of predicted ratings was 11.68. These findings suggest the model maintained stable accuracy across structurally diverse vehicles, with minimal perceptual drift even under varying door-opening dynamics.

Attribute-wise (Figure 5.13), the lowest standard deviation of user ratings was seen for Luxurious–Cheap (13.47), indicating relatively consistent perceptions among participants. However, Easy-to-pull–Hard-to-pull exhibited the highest user rating variance (STD = 20.58 actual, 19.41 predicted), likely due to individual differences in assessing physical effort. Other attributes such as Damped–Recoiling and New–Old showed moderate dispersion, but predictions still tracked actual scores closely, reflecting the model's ability to generalize across both physical and emotional descriptors.

Overall, these analyses confirm that the CNN-based model can robustly map mechanical force profiles to perceptual impressions, with prediction variance well contained within perceptual uncertainty. The consistency across cars and attributes underscores the model's generalizability and reliability.

## 5.6 Perception-to-Force Generation

This section presents a model designed to generate force profiles from perceptual attribute inputs. The architecture builds on a pretrained decoder originally trained to infer perceptual ratings from force signals. By reversing the direction of inference, this model enables synthesis of mechanical signals that reflect user-defined perceptual intent.

### 5.6.1 Proposed Architecture

The model follows an encoder–decoder structure where the input is a 7-dimensional perceptual attribute vector. This vector is reshaped to size $(7, 1)$ and passed through a series of one-dimensional convolutional layers to produce a latent representation compatible with the pretrained decoder. The encoder consists of four convolutional layers: 64 filters with kernel size 2, followed by 128, 256, and 512 filters with kernel size 3. A max pooling layer is applied after the first convolution to reduce temporal resolution. The output is flattened and projected through a dense layer, then reshaped to match the decoder's expected latent shape of $(155, 128)$.

The decoder reconstructs the full-resolution force signal through two residual convolutional blocks. The first block applies two convolutional layers with 64 filters and kernel size 3, followed by upsampling from 155 to 310 time steps. The second block uses two convolutional layers with 128 filters and kernel size 3, followed by another upsampling stage to reach 620 time steps. A final dense layer with linear activation outputs a 621-point force profile.

Decoder weights are transferred from the pretrained signal-to-rating model. These weights are fixed during training and not updated. Only the encoder and final dense layer are trainable. This reuse ensures consistency with the signal structure previously learned from perceptual ratings and reduces the total number of trainable parameters. The model is implemented in TensorFlow and compiled as a single, end-to-end architecture.

#### 5.6.1.1 Model Training Procedure

The network is trained to generate normalized force profiles of 621 time steps from 7-dimensional perceptual inputs. All inputs and outputs are scaled to the range $[0, 100]$ prior to training. The model is optimized using the Mean Squared Error (MSE) loss function and the Adam optimizer

with a learning rate of $10^{-4}$ and a batch size of 8. Early stopping is applied based on validation loss.

To ensure generalization across unseen data and to assess performance robustness, 5-fold cross-validation is used. The dataset, which includes both original and augmented car profiles, is partitioned so that each fold includes a balanced distribution across car types and their variations. For each fold, the model is trained on four subsets and validated on the remaining one. Evaluation metrics include Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE).

The decoder is not fine-tuned during this process. It retains the structure and weights learned from the signal-to-rating task, ensuring that force synthesis remains aligned with human-rated perceptual patterns.

### 5.6.2 Numerical Evaluation

#### 5.6.2.1 Dataset and Evaluation Protocol

The dataset preparation, normalization, and augmentation follow the same strategy used in the force-to-perception model described earlier in Section 5.3.4. Each sample comprises a 7-dimensional perceptual rating vector paired with a corresponding normalized force profile consisting of 621 time steps. Both original and augmented profiles are included, resulting in a total of 545 samples across six car models. All values are scaled to the range $[0, 100]$.

To ensure robust evaluation, a 5-fold cross-validation scheme is used. The data is split such that each fold contains a balanced distribution of vehicle types and their augmentation variants. The model is trained and validated on disjoint folds, and the results are averaged across all five iterations.

#### 5.6.2.2 Error Metrics

Model performance is assessed using two standard regression metrics. The Mean Absolute Error (MAE) is defined as

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i|,$$

and the Root Mean Square Error (RMSE) is defined as

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2},$$

where $y_i$ and $\hat{y}_i$ are the ground truth and predicted force values, respectively, and $N = 621$ is the signal length. These metrics quantify how closely the generated signals match the original human-rated force profiles in both magnitude and variation.

Predicted and actual waveforms are directly compared to preserve temporal continuity and to evaluate the signal-level accuracy of the generation process. This approach reflects the practical requirement for waveform-level consistency in haptic signal rendering, beyond aggregate statistical trends.

#### 5.6.2.3 Results and Analysis

Table 5.6 summarizes the MAE and RMSE values obtained across the six vehicle categories. Each entry reflects the mean and standard deviation computed across all folds and all profile variants.

**Table 5.6:** Prediction error summary for each car model. Results are reported as mean ± standard deviation.

| Car Model | MAE (%) | RMSE (%) |
|---|---|---|
| BC3 | $0.546 \pm 0.308$ | $0.695 \pm 0.351$ |
| FORD | $0.567 \pm 0.309$ | $0.740 \pm 0.368$ |
| G90 | $0.620 \pm 0.437$ | $0.785 \pm 0.495$ |
| K3 | $0.537 \pm 0.302$ | $0.654 \pm 0.329$ |
| K8 | $0.623 \pm 0.416$ | $0.761 \pm 0.451$ |
| MQ4 | $0.577 \pm 0.319$ | $0.729 \pm 0.369$ |

The best performance is observed for the K3 profiles, which yield the lowest MAE (0.537) and RMSE (0.654). This indicates that the generated waveforms for K3 most closely match the original force signals. In contrast, G90 and K8 show the highest RMSE (0.785 and 0.761, respectively), possibly due to more complex force dynamics or irregular signal transitions that are harder to approximate.

Figure 5.14 shows a violin plot of MAE distributions across cars. This plot complements the tabulated summary by visualizing the spread and density of errors for each vehicle class. Despite

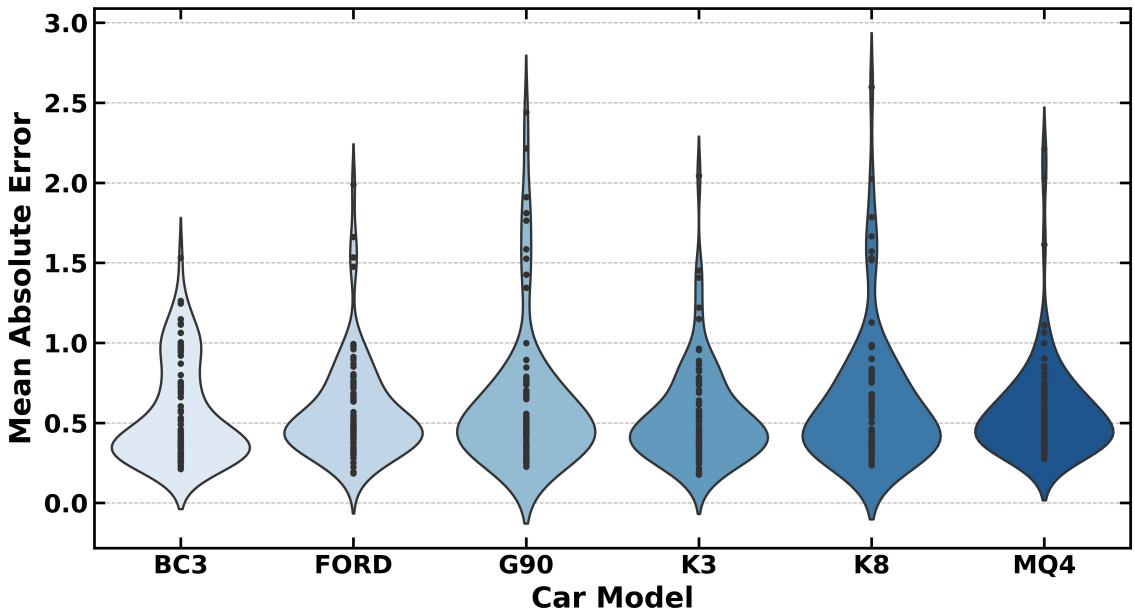


**Figure 5.14:** Distribution of MAE across car models in the perception-to-force task. Each violin shows the density and spread of errors within one car class.

minor variations, all distributions remain tightly bounded, with most samples centered around the mean. The narrow error range confirms that the model maintains consistent predictive accuracy across both real and augmented profiles.

Overall, the perception-to-force model demonstrates robust reconstruction of mechanical force profiles from abstract perceptual inputs. The use of a pretrained decoder contributes to stable generation across diverse signal types, while the learned encoder adapts perceptual ratings into physically grounded output trajectories.

## 5.7 Perceptual Evaluation

This section presents the perceptual evaluation of the proposed perception-to-force generation model. Two experiments were conducted. The first assessed whether force profiles generated from individual perceptual attributes matched user expectations. The second examined the usability and real-time effectiveness of the authoring interface when operated by participants.

### 5.7.1 Experiment 1: Attribute-Based Perceptual Evaluation

This experiment was designed to evaluate whether the force profiles generated by the perception-to-force model accurately reflected the intended perceptual attributes. Since the model accepts a seven-dimensional perceptual input vector and produces a corresponding kinesthetic force profile, the primary goal was to assess if the output profile conveyed the targeted attribute to users during physical interaction.

To perform this evaluation, a perceptual matching approach was selected. Participants were asked to experience a rendered force signal and then provide ratings across all seven attribute dimensions using continuous sliders. This method was chosen over alternatives such as binary forced-choice or rank-ordering tasks. Forced-choice techniques are limited to single-attribute comparisons [120], and rank-ordering lacks the resolution required for accurate quantitative evaluation. In contrast, perceptual matching yields continuous, multidimensional feedback aligned with the structure of the model's input and enables direct comparison between intended and perceived values [121, 122].

#### 5.7.1.1 Force Generation Interface and Stimuli

A custom graphical user interface was developed to enable real-time generation of force profiles from perceptual inputs (see Figure 5.15). The interface was designed to support intuitive control over the seven bipolar adjective pairs used in this study: Easy-to-pull – Hard-to-pull, Damped – Recoiling, Consistent – Stepwise, Pleasant – Unpleasant, New – Old, Comfort – Discomfort, and Luxurious – Cheap. Each dimension was mapped to a vertical slider ranging from 0 to 100, forming a seven-dimensional input vector that could be continuously adjusted. This input vector was streamed into the trained perception-to-force model to synthesize a 621-sample kinesthetic force signal. The interface also includes additional controls for signal smoothing, playback, export, and saving. A brief instruction panel explained the function of each component.

For user interaction during the experiment, the stimuli were generated in advance using the developed interface. Each profile was created by adjusting a single slider to either 10 or 90, while keeping all other sliders fixed at the neutral midpoint of 50. The extreme values of 10 and 90 were used instead of 0 and 100 to avoid potential boundary effects in the model response and to ensure

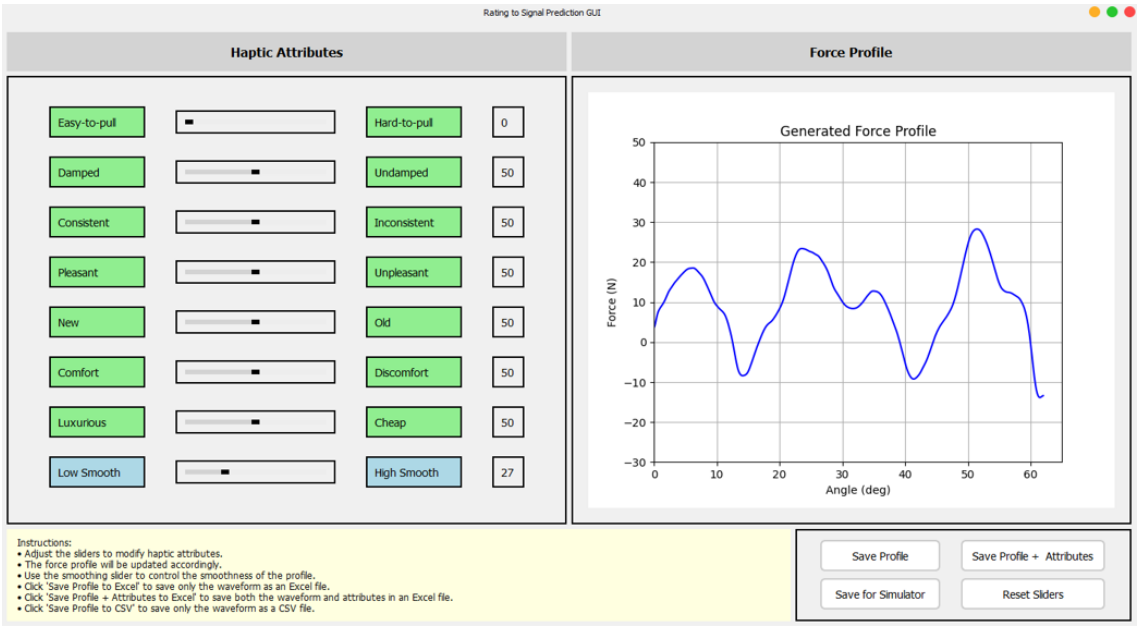


**Figure 5.15:** Graphical user interface for real-time perceptual-to-force synthesis. In Experiment 1, the interface was operated by the experimenter to generate stimulus profiles. In Experiment 2, it was used directly by participants for interactive force design.

stability in signal generation. This procedure allowed each attribute to be tested independently without interference from other dimensions. The resulting force profiles were saved as individual time-series signals and later loaded into the programmable car door simulator for playback during the experiment. A total of fourteen stimuli were prepared in this way, covering both ends of all seven perceptual dimensions.

#### 5.7.1.2 Participants and Procedure

Sixteen participants (14 male, 2 female; ages 24–52) were recruited for the study. All reported normal tactile perception and no known sensory impairments. During each trial, one of the 14 generated force profiles was presented in randomized order. Participants physically interacted with the force signal using the car door simulator (see Section. 5.3.2).

To eliminate non-haptic cues, participants wore noise-canceling headphones and eye masks throughout the experiment. After each interaction, they rated their perception using seven continuous sliders (0–100 scale), one for each attribute.

Perceptual accuracy was quantified by computing the absolute difference between the original target value and the participant's rating on the manipulated attribute. These errors were used to evaluate the alignment between the intended perceptual input and the perceived response.

#### 5.7.1.3 Results and Analysis

Figure 5.16 shows the distribution of attribute rating scores provided by participants, along with the corresponding mean ratings for each profile type. Across most attributes, participants' ratings were closely aligned with the intended perceptual targets, indicating that the proposed system effectively synthesized force profiles that conveyed the desired perceptual qualities. This outcome highlights the model's ability to generate meaningful kinesthetic variations that are interpretable and distinguishable to users.

This effect was most pronounced for physically grounded attributes. Among these, Easy-to-pull – Hard-to-pull exhibited the lowest mean absolute error (5.83), followed by Damped – Recoiling (7.08) and Consistent – Stepwise (6.67). These results confirm that the kinesthetic cues delivered by the system were effective in communicating mechanical characteristics of the door interaction. A closer look at the profile-specific errors reveals further insights: for Easy-to-pull – Hard-to-pull, the Min profile produced a slightly higher error (6.67) compared to the Max profile (5.00), indicating a minor overestimation bias for low-effort cues. Conversely, Damped – Recoiling showed increased error in the Max condition (8.33), suggesting that high-intensity recoiling cues may be more difficult to interpret consistently.

Attributes that lie at the boundary between physical and emotional perception exhibited more varied outcomes. Comfort – Discomfort, for instance, demonstrated moderate accuracy with a mean error of 9.17. However, the difference between Min (12.50) and Max (5.83) profile errors was substantial, implying that participants found it easier to detect and agree upon sensations associated with discomfort rather than comfort. This asymmetry may indicate that negative affective impressions evoke stronger or more consistent perceptual responses than positive ones under kinesthetic rendering.

In contrast to physical descriptors, emotional attributes exhibited greater perceptual variation across participants. Pleasant – Unpleasant showed moderate agreement, with a mean absolute

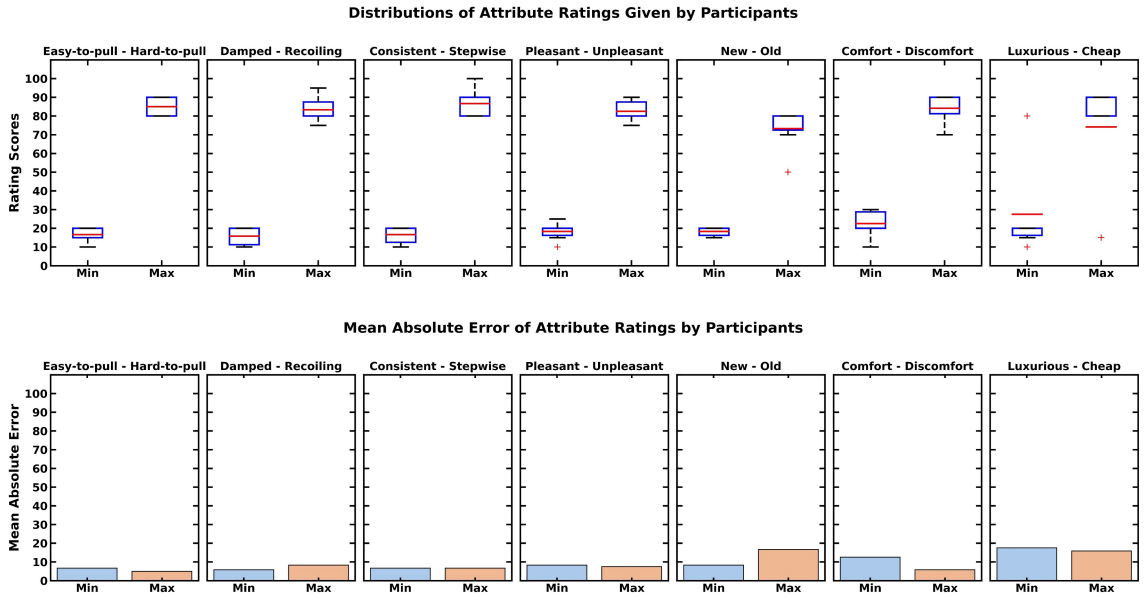


**Figure 5.16:** Results from the perceptual study evaluating attribute-wise performance. For each attribute, two profiles were generated representing opposite extremes (e.g., 10 as New and 90 as Old), while keeping other attributes at a neutral (50) level. The top plots show the distribution of participant ratings, and the bottom plots present the corresponding mean ratings for each profile.

error of 7.92, while New – Old (12.50) and Luxurious – Cheap (16.67) showed higher deviation. Luxurious – Cheap produced the largest standard deviation (26.49), reflecting greater spread in participant responses. Although most ratings were numerically near intended targets, variability in qualitative interpretation was evident. For example, a profile designed to convey "luxury" was occasionally perceived as "cheap," suggesting variation in subjective judgments rather than a flaw in signal generation. Similarly, the Max profile for New–Old showed high error (16.67), indicating that some abstract descriptors may be less consistently interpreted through kinesthetic feedback.

When analyzed by category, physical attributes consistently yielded lower error (mean: 7.19) compared to emotional attributes (mean: 12.36). This confirms that the proposed system effectively conveys concrete mechanical characteristics through kinesthetic modulation. The overall mean absolute error across all conditions was 9.40, with individual profile types contributing differently depending on the attribute. These findings highlight the system's strength in reliably encoding physical qualities while also demonstrating its potential for more abstract attributes, which may involve broader perceptual interpretations.

### 5.7.2 Experiment 2: Real-Time Interaction and Usability Study

The second experiment evaluated the usability and perceptual controllability of the system in a real-time interactive setting. Unlike Experiment 1, where stimuli were predefined, this study allowed users to freely configure perceptual attributes of their choice. Participants adjusted the perceptual sliders, generated the corresponding force profile through the trained model, and immediately experienced the resulting feedback on the car door simulator. The overall interaction process is illustrated in Figure 5.17.

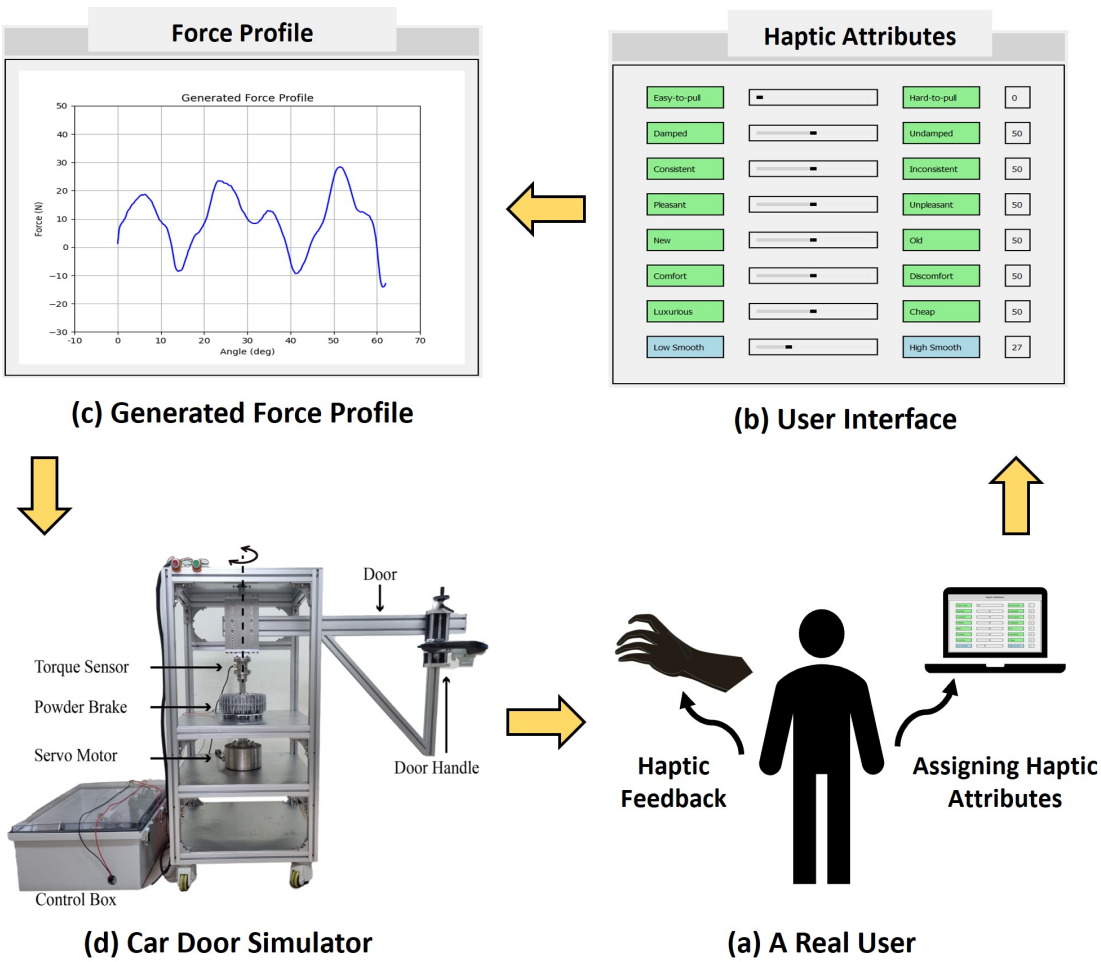


**Figure 5.17:** Overview of the real-time interaction process used in the usability study. Participants adjusted perceptual sliders, generated corresponding force profiles, and experienced the output through the car door simulator.

#### 5.7.2.1 Interface and Procedure

The same graphical interface used in Experiment 1 was operated directly by the participants. Each user manipulated the seven perceptual sliders to define their intended configuration. The resulting 7D vector was streamed to the trained perception-to-force model at 20 Hz, which generated a corresponding 621-sample force signal. This signal was displayed onscreen and rendered through the car door simulator in real time.

Before starting the main task, participants completed a short familiarization session to explore the interface and understand how their inputs affected the output. During the main interaction phase, they were allowed to freely adjust the sliders and evaluate the resulting haptic sensations. No restrictions were placed on the number of trials or the sequence of adjustments. Following the interaction phase, each participant completed a five-item questionnaire designed to assess system fidelity and usability.

**System Fidelity:**

- Correctness: The rendered feedback matched the intended perceptual input.
- Differentiability: Changes in sliders produced clearly different outputs.
- Realism: Physical feedback felt plausible and mechanically valid.

**Usability:**

- Ease of Use: Interface was simple and responsive.
- Learnability: The mapping between sliders and force output was intuitive.

#### 5.7.2.2 Results and Analysis

Each participant provided five ratings at the end of the study, corresponding to questionnaire items targeting perceptual correctness, realism, differentiability, ease of use, and learnability. With 16 participants, a total of 80 responses were recorded. All ratings were collected on a 7-point Likert scale (1 = strongly disagree, 7 = strongly agree).

Figure 5.18 presents the distribution of ratings across the five dimensions. Among the measures, *Differentiability* received the highest average rating (mean = 5.94), followed by *Ease of Use*

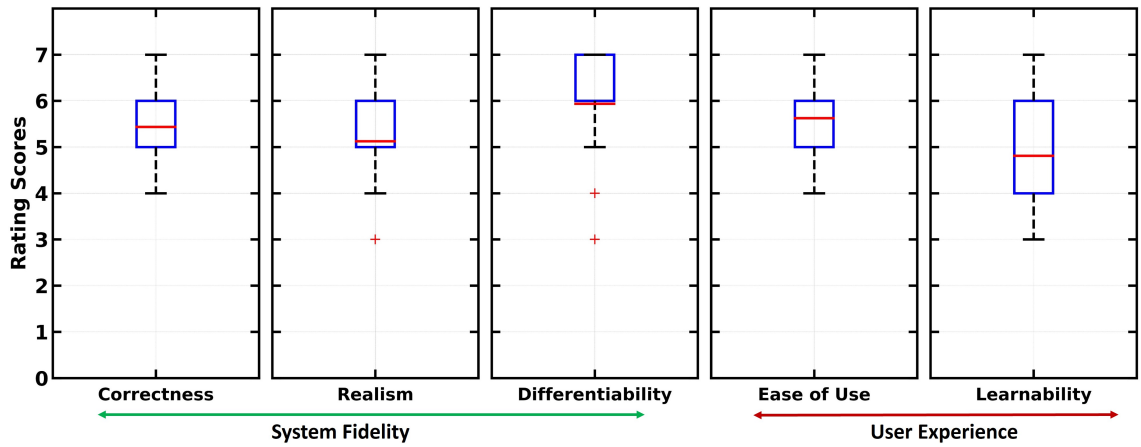


**Figure 5.18:** Rating scores from the user experience study. Boxplots show distribution across all participants for each measure.

(mean = 5.63) and *Correctness* (mean = 5.44). These results indicate that participants generally found the system intuitive and capable of producing clearly distinguishable force responses that aligned with their intended perceptual adjustments.

Ratings for *Realism* (mean = 5.13) and *Learnability* (mean = 4.81) were relatively lower. While still above the neutral point, the reduced score for Learnability suggests a moderate learning effort was required for some users to fully understand the effect of each slider. In particular, one participant (User 4) consistently gave lower ratings across all items, which may reflect individual difficulty in mapping subjective perception to force-based feedback or a stricter internal standard.

Further analysis of qualitative behavior during the interaction phase revealed that some participants encountered semantic inconsistencies when adjusting perceptual sliders. For example, increasing “Easy-to-pull” to its maximum sometimes produced profiles that felt mechanically unrealistic or unpleasant, leading to dissonance between semantic intent and physical sensation. Such mismatches may have contributed to lower Learnability scores, as participants needed to reconcile linguistic descriptors with haptic outcomes.

Overall, these findings demonstrate that the proposed perception-to-force system delivers perceptually valid and controllable force profiles. The generally high ratings, supported by low interquartile spreads in several measures, confirm that the interface performs robustly in real-time authoring scenarios.

## 5.8 Discussion

This study presented a bidirectional framework for understanding and generating the perceptual experience of opening a car door. The system successfully linked measurable physical signals to subjective user impressions by predicting perceptual ratings from force profiles and generating force profiles from perceptual input. The model performance, supported by both numerical metrics and perceptual experiments, indicates that car door haptics can be reliably modeled and reproduced using deep learning methods. The results suggest practical applications in perception-aware design, simulation, and rapid prototyping of car doors.

### 5.8.1 Mapping Physical Force to Perception

The core contribution of this work lies in establishing a connection between the physical signal space and the cognitive perception space. The physical space consisted of normalized torque profiles representing the dynamics of real car doors, while the perception space was derived from user ratings along semantically meaningful adjective pairs. The CNN-based prediction model demonstrated that the shape, slope, and amplitude of force profiles encode sufficient information to infer how users perceive the mechanical interaction.

Among the seven adjective pairs evaluated, attributes such as ”Comfort–Discomfort” and ”Damped–Recoiling” showed high predictive reliability. These results indicate that human perception of mechanical realism and smoothness is tightly coupled with measurable physical features. The model achieved an average MAE of less than 3.5 percent for most attributes, confirming that small variations in force shape can produce distinguishable changes in perception.

### 5.8.2 Understanding Prediction Error Patterns

Prediction errors remained within acceptable perceptual bounds, as determined by the observed standard deviation of user responses. The average standard deviation across participants was approximately 21 units on a normalized 100-point scale, while the model's average prediction error remained below 10 units. This level of error is consistent with perceptual uncertainty and is unlikely to affect the user's impression of the system.

The reverse model, trained to generate force profiles from perceptual input, also showed strong numerical performance, with an average MAE below 0.65 percent. Despite this low error, some profiles, particularly those associated with abstract descriptors such as ”Luxurious”, were more difficult to reconstruct consistently. These deviations are likely influenced by subjective factors that go beyond mechanical feedback, including prior expectations or associations with brand or product class.

### 5.8.3 Insights from Perception-to-Force Evaluation Study

This section evaluates the perceptual validity and interactive usability of the proposed perception-to-force framework through two complementary experiments. While offline regression metrics indicated strong model performance, the goal here was to determine whether users could meaningfully interpret and control the generated force profiles in both constrained and open-ended settings.

In Experiment 1, participants rated force profiles generated by manipulating individual perceptual attributes while keeping all others fixed at neutral. Results confirmed that the model could effectively convey concrete physical properties such as Easy-to-pull, Damped – Recoiling, and Consistent – Stepwise, all of which showed low perceptual error and consistent interpretation across participants.

In contrast, emotional and abstract descriptors led to greater variability. Notably, for Luxurious – Cheap, some users rated the same profile as “luxurious” while others rated it as “cheap,” producing a bidirectional distribution despite numeric closeness to the target. This pattern reflects the inherently subjective nature of such attributes, where personal preferences, prior experiences, and cultural associations significantly influence interpretation. Similar inconsistencies, though less pronounced, were also observed for Pleasant – Unpleasant and New – Old, suggesting that perceptual misalignment can arise from semantic ambiguity rather than system error alone.Some variability may be due to the decision to manipulate one attribute at a time while keeping others fixed at neutral. In practice, many perceptual qualities are closely related; for example, easy-to-pull often aligns with comfortable. Holding other sliders at midpoints may have limited these natural associations, especially for emotional descriptors. Although this design helped isolate the

effect of each attribute, allowing multiple attributes to vary would make the study highly complex and reduce clarity, making it difficult to understand which factor influenced perception.

Experiment 2 extended the evaluation to real-time interaction. Participants directly manipulated the perceptual sliders and received immediate haptic feedback. High ratings for differentiability, correctness, and ease of use indicated that users generally found the system responsive and intuitive. However, relatively lower scores for realism and learnability pointed to challenges in interpreting certain sliders, particularly those representing abstract qualities.

Additionally, extreme settings on some sliders occasionally generated force profiles that felt mechanically implausible or lacked coherent structure. For instance, maximal Easy-to-pull values sometimes produced signals devoid of realistic resistance. Such issues may arise from nonlinear mappings in the model or perceptual saturation effects at the input extremes.

To address these limitations, future designs should consider integrating calibrated reference profiles and concise explanations for each attribute, enabling participants to build a more grounded expectation of how perceptual descriptors map to physical sensations. Attribute-specific scaling strategies could also help ensure that extreme values yield perceptually meaningful and mechanically valid outputs.

In summary, the system demonstrated strong performance in generating interpretable and controllable force feedback for physically grounded attributes. At the same time, the findings highlight the difficulty of representing emotional or semantic dimensions through kinesthetic feedback alone, emphasizing the need for improved semantic grounding and user-centered design in perceptual interfaces.

### 5.8.4 Challenges in Interpreting Perceptual Ratings of Haptic Attributes

This study is fundamentally centered on user perception, as the effectiveness of the proposed framework depends on how participants interpret and rate force-based haptic feedback. Since perceptual descriptors serve as both model inputs and evaluation targets, understanding how consistently users interpret these terms is essential. The perceptual space defined in this work consisted of seven bipolar adjective pairs selected to reflect both physical and emotional qualities of the car door interaction. These attributes naturally fall into two broad categories: physical and emotional.

Each category brings its own set of challenges for signal generation and user interpretation.

Physical attributes, including Easy to pull, Damped versus Recoiling, and Consistent versus Stepwise, are closely tied to mechanical sensations that users can experience directly. These are generally more concrete and easier to map to kinesthetic signals. However, they are still influenced by individual differences. For example, ratings for Easy to pull were sometimes affected by the user's physical strength or prior experience with different vehicle types. In some cases, lower resistance was not seen as a positive feature but was interpreted as offering insufficient mechanical feedback. This divergence in perception was especially notable across gender, where different expectations about effort and mechanical feedback appeared to shape how users evaluated the same force profile.

In contrast, emotional attributes such as Pleasant versus Unpleasant, Luxury versus Cheap, and New versus Old introduce greater perceptual ambiguity. These dimensions are semantically richer and rely more on subjective interpretation, making it harder to associate them with a specific physical sensation. For instance, while some users interpret a smooth and lightweight door as luxurious, others perceive the same sensation as lacking substance. These differences stem from personal experience, product expectations, and cultural context, all of which contribute to higher inter-user variability in emotional ratings.

Although such perceptual inconsistencies are expected in user studies involving semantic descriptors, several measures were taken to reduce their impact. Before the experiment, participants received clear written instructions about the task and the meaning of each attribute. Each slider was accompanied by a short sentence explaining its interpretation in the context of car door feedback. Participants wore eye masks and noise-canceling headphones to remove visual or auditory influence and were instructed to focus only on the mechanical sensation. Even with these precautions, complete consensus could not be achieved across all participants. Nevertheless, by normalizing the ratings within each user and analyzing trends across the group, the study was able to reveal consistent patterns for most physical attributes, and meaningful tendencies even in the emotional ones.

These findings suggest that while perception-based modeling is a powerful tool for intuitive force generation, care must be taken in both the design and evaluation of perceptual interfaces.

Mechanical descriptors may benefit from strength-aware calibration, while emotional ones may require anchoring or contextualization to reduce interpretation gaps. Designing future systems that better support user alignment across these categories will be critical for achieving consistent and expressive haptic experiences.

### 5.8.5 Implications for Perception-Centered Design

Perception-based modeling offers a structured and explainable approach to designing haptic signals. Rather than relying on discrete labels or low-level physical parameters, this framework accepts continuous perceptual inputs that reflect how users naturally describe physical interactions. In practice, users are more likely to define a mechanical experience by stating that a car door feels "easy to pull" or "recoiling" than by referencing exact force magnitudes. Perception-driven modeling supports this formulation by enabling intuitive control and allowing designers to generate force feedback that aligns with subjective experience.

This approach supports both forward and inverse applications. Perceptual ratings can be predicted from measured signals to evaluate user experience. Conversely, desired perceptual attributes can be used to generate corresponding force profiles. These capabilities make the framework useful not only for perceptual evaluation but also for real-time design and prototyping.

The core concept is based on constructing two well-defined spaces: a perceptual space composed of human-interpretable haptic attributes, and a physical space consisting of measured or synthesized haptic signals. For effective mapping between these spaces, the chosen attributes must be semantically meaningful, perceptually grounded, and consistent across users. The signal representations must also reflect the physical characteristics of the interaction modality, whether through force, vibration, or pressure.

Although this work focused on kinesthetic interaction through force profiles, the core concept of perception-based modeling involves constructing a mapping between a perceptual space and a physical signal space. The perceptual space should be defined using haptic attributes that are semantically meaningful and interpretable by users. These attributes must be identified carefully based on perceptual relevance, consistency across users, and coverage of the intended interaction domain. In parallel, the physical space consists of measured or synthesized haptic signals such as

force trajectories, acceleration waveforms, or pressure patterns.

Each haptic modality introduces distinct signal characteristics and perceptual encoding challenges. For example, tactile textures are represented by high-frequency acceleration signals typically in the range of 20 to 1000 Hz, capturing fine surface vibrations generated during sliding contact. These differ significantly from the low-frequency force signals used in kinesthetic feedback, which generally fall below 100 Hz and reflect large-scale resistance and damping. As a result, the model architecture and training formulation used in this work are not directly applicable to tactile or vibrotactile rendering tasks.

To adapt this approach to other haptic properties, both the perceptual descriptors and the signal representation must be redesigned. Texture rendering, alert vibrations, and dynamic pressure-based cues each require modality-specific signal structures, appropriate sampling strategies, and tailored perceptual dimensions. While the current framework demonstrates how perceptual attributes can drive force signal generation, generalizing it to other domains requires careful construction of the perceptual and physical spaces based on the nature of the target haptic interaction.

Perception-based haptic modeling enables interpretable and controllable signal generation, but it must be adapted to the specific signal properties and perceptual structure of each modality. Designing such systems requires precise attribute definition, awareness of user variability, and alignment with the unique characteristics of the target feedback type.

## 5.9 Conclusion

This work introduced a bidirectional framework for modeling and synthesizing perceptual impressions of car door interactions based on mechanical force profiles. By integrating sensor-based signal capture, user-driven perceptual data, and deep neural architectures, the system enables the prediction of perceptual ratings from physical input and the generation of force profiles from user-defined perceptual intent.

In the forward direction, a CNN-based model was trained to estimate user ratings across seven bipolar perceptual attributes. The model achieved a mean absolute error below 3.5 percent, confirming that force signals encode salient features that can be reliably decoded through learned temporal and spatial representations. Attribute-level and car-wise evaluations demonstrated con-

sistent performance across multiple interaction conditions.

In the reverse direction, a perception-to-force model generated kinesthetic signals from perceptual vectors with low reconstruction error. Two perceptual studies were conducted to validate the synthesized output. In the first study, participants rated force profiles generated for individual attributes. The results showed strong alignment with intended perceptual targets, particularly for physically grounded dimensions such as ease, damping, and consistency. Higher variability was observed for emotional attributes such as luxuriousness, reflecting the challenge of conveying abstract qualities through kinesthetic feedback alone. In the second study, participants used a real-time authoring interface to define perceptual intent and evaluate the corresponding force output. User ratings confirmed that the system enabled controllable and intuitive feedback generation, though some participants reported difficulty mapping subjective descriptors to physical sensations, particularly for less tangible attributes.

Together, these findings demonstrate that both perceptual inference and synthesis of car door haptics can be achieved through a unified, data-driven approach. The proposed system offers a foundation for perceptually informed interaction design, reducing reliance on physical prototyping and enabling real-time authoring tools tailored to user intent. Future work may explore multimodal extensions, real-time personalization, and broader applications in mechanical and automotive interface design.

# Chapter 6

# Conclusion and Future Work

## 6.1 Conclusion

This thesis presented a comprehensive framework for modeling, predicting, and rendering human haptic perception across both tactile and kinesthetic domains. Motivated by the growing demand for perceptually intelligent systems in virtual and remote interaction, the work explored how subjective sensations can be computationally understood and recreated from physical signals and user-defined perceptual goals.

In the tactile domain, the thesis introduced a perceptual modeling pipeline for surface textures. A perceptual attribute space was constructed through psychophysical studies using real textured materials, identifying four key dimensions: rough–smooth, flat–bumpy, sticky–slippery, and hard–soft. A multimodal deep learning model was trained to predict perceptual ratings from a combination of visual and tactile sensor data. This model enabled automatic perceptual labeling of textures, supporting intuitive tagging and content retrieval. Additionally, a Fourier-enhanced Transformer Encoder Network was proposed to synthesize high-frequency tactile signals from interaction parameters such as speed and force. This model provided efficient and high-fidelity rendering suitable for real-time applications, validated through both quantitative metrics and perceptual testing.

In the kinesthetic domain, the thesis focused on the perception of car door interaction. A bidirectional modeling approach was proposed to map between force profiles and user-defined perceptual attributes. Initially, a residual CNN-based model was trained to predict perceptual qualities from measured torque signals, using a perceptual space derived from user studies with both real and simulated car doors. Subsequently, an inverse decoder-based model was introduced to generate force profiles from target adjectives such as 'Easy-to-Pull' or 'luxurious.' Together,

these models support the perceptual evaluation and design of mechanical feedback systems without manual tuning or physical prototyping, offering a scalable solution for user-centric product interaction.

Across all domains explored, this thesis emphasized a bidirectional connection between physical interaction signals and user cognition, showing that haptic perception can not only be predicted but also generated in a controllable and scalable manner. The methods proposed here provide foundational steps toward the development of haptic systems that are aware of, and adaptable to, human perception.

## 6.2 Future Directions

While the methods presented in this thesis demonstrate promising results, several opportunities remain for extending this work. One major direction involves improving model generalization across broader interaction contexts. Both tactile and kinesthetic datasets can be expanded to include a wider variety of materials, motion paths, and usage conditions to better reflect real-world variability.

Another area of interest is the integration of higher-level semantic understanding of perception. While this thesis focused on physical descriptors such as roughness or smoothness, future work could incorporate contextual or emotional dimensions of haptics, enabling systems that respond not just to texture properties but also to user intent and task demands.

Additionally, the perceptual attribute prediction and generation frameworks developed here can be applied to other interactive devices, such as pens, tools, gloves, or consumer products. Embedding such models in real-time rendering systems would allow for dynamic, user-specific feedback in teleoperation, training, or immersive environments.

Finally, there is potential for developing closed-loop systems that combine sensing, perception modeling, and feedback rendering in a unified architecture. Such systems could adapt haptic output based on real-time user response, enabling more personalized and immersive tactile interactions.

Through these directions, the foundational work of this thesis can evolve into practical systems that advance the fidelity, intelligence, and user-centered design of haptic technologies.

# Bibliography


[1] T. Yoshioka, S. J. Bensmaia, 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 research*, vol. 24, no. 1-2, pp. 53–70, 2007.

[2] K. Yoshida *et al.*, "The dimensions of tactile perception of surfaces," *Journal of Texture Studies*, vol. 12, pp. 123–135, 1968.

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

[4] 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 research*, vol. 22, no. 3, pp. 127–140, 2005.

[5] R. H. LaMotte, "Softness discrimination with a tool," *Journal of neurophysiology*, vol. 83, no. 4, pp. 1777–1786, 2000.

[6] H. Culbertson and K. J. Kuchenbecker, "Ungrounded haptic augmented reality system for displaying roughness and friction," *IEEE/ASME Transactions on Mechatronics*, pp. 1839–1849, 2017.

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

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

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

[10] J. Wu, N. Li, W. Liu, G. Song, and J. Zhang, "Experimental study on the perception characteristics of haptic texture by multidimensional scaling," *IEEE transactions on haptics*, pp. 410–420, 2015.

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

[12] W. Hassan, A. Abdulali, and S. Jeon, "Authoring new haptic textures based on interpolation of real textures in affective space," *IEEE transactions on industrial electronics*, pp. 667–676, 2019.

[13] M. Ibrahim Awan and S. Jeon, "Estimating perceptual attributes of haptic textures using visuo-tactile data," *IEEE Access*, vol. 13, pp. 109 931–109 945, 2025.

[14] W. Hassan, J. B. Joolee, and S. Jeon, "Establishing haptic texture attribute space and predicting haptic attributes from image features using 1d-cnn," *Scientific Reports*, vol. 13, no. 1, p. 11684, 2023.

[15] M. Strese, C. Schuwerk, and E. Steinbach, "Surface classification using acceleration signals recorded during human freehand movement," in *2015 IEEE World Haptics Conference (WHC)*. IEEE, 2015.

[16] 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 *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023.

[17] S. Lu, M. Zheng, M. C. Fontaine, S. Nikolaidis, and H. Culbertson, "Preference-driven texture modeling through interactive generation and search," *IEEE transactions on haptics*, pp. 508–520, 2022.

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

[19] 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 *Proceedings of the 29th ACM Symposium on Virtual Reality Software and Technology*, 2023, pp. 1–9.

[20] 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 Transactions on Haptics*, 2024.

[21] F. Yang, C. Feng, Z. Chen, H. Park, D. Wang, Y. Dou, Z. Zeng, X. Chen, R. Gangopadhyay, A. Owens *et al.*, "Binding touch to everything: Learning unified multimodal tactile representations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 26 340–26 353.

[22] M. Strese, C. Schuwerk, A. Iepure, and E. Steinbach, "Multimodal feature-based surface material classification," *IEEE transactions on haptics*, vol. 10, no. 2, pp. 226–239, 2016.

[23] 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 in Computer Science*, vol. 6, p. 1401560, 2024.

[24] H. Li and H. Zhang, "Classification method of visual-tactile fusion dataset based on cnn-tcn," in *2023 8th International Conference on Control, Robotics and Cybernetics (CRC)*. IEEE, 2024, pp. 295–300.

[25] W. Byeon, M. Liwicki, and T. M. Breuel, "Texture classification using 2d lstm networks," in *2014 22nd international conference on pattern recognition*. IEEE, 2014, pp. 1144–1149.

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

[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[28] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," *Advances in neural information processing systems*, vol. 28, 2015.

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

[30] I. Bounoua, Y. Saidi, R. Yaagoubi, and M. Bouziani, "Deep learning approaches for water stress forecasting in arboriculture using time series of remote sensing images: Comparative study between convlstm and cnn-lstm models," *Technologies*, vol. 12, no. 6, p. 77, 2024.

[31] S. Yao, Y. He, L. Zhang, W. Yang, Y. Chen, Q. Sun, Z. Zhao, and S. Cao, "A convlstm neural network model for spatiotemporal prediction of mining area surface deformation based on sbas-insar monitoring data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–22, 2023.

[32] S. Shin and S. Choi, "Geometry-based haptic texture modeling and rendering using photometric stereo," in *2018 IEEE Haptics Symposium (HAPTICS)*. IEEE, 2018, pp. 262–269.

[33] L. Kim, A. Kyrikou, G. S. Sukhatme, and M. Desbrun, "An implicit-based haptic rendering technique," in *IEEE/RSJ international conference on intelligent robots and systems*, vol. 3. IEEE, 2002, pp. 2943–2948.

[34] N. Zafer, "Constraint-based haptic rendering of a parametric surface," *Proceedings of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering*, vol. 221, no. 3, pp. 507–517, 2007.

[35] A. M. Okamura, K. J. Kuchenbecker, and M. Mahvash, "Measurement-based modeling for haptic rendering," *Haptic Rendering: Algorithms and Applications*, pp. 443–467, 2008.

[36] J. M. Romano and K. J. Kuchenbecker, "Creating realistic virtual textures from contact acceleration data," *IEEE Transactions on haptics*, vol. 5, no. 2, pp. 109–119, 2011.

[37] A. Abdulali and S. Jeon, "Data-driven rendering of anisotropic haptic textures," in *International AsiaHaptics Conference*. Springer, 2016, pp. 401–407.

[38] Y. Ujitoko and Y. Ban, "Vibrotactile signal generation from texture images or attributes using generative adversarial network," in *Haptics: Science, Technology, and Applications: 11th International Conference, EuroHaptics 2018, Pisa, Italy, June 13-16, 2018, Proceedings, Part II 11*. Springer, 2018, pp. 25–36.

[39] S. Shin, R. H. Osgouei, K.-D. Kim, and S. Choi, "Data-driven modeling of isotropic haptic textures using frequency-decomposed neural networks," in *2015 IEEE World Haptics Conference (WHC)*, 2015.

[40] J. B. Joolee and S. Jeon, "Data-driven haptic texture modeling and rendering based on deep spatio-temporal networks," *IEEE Transactions on Haptics*, vol. 15, no. 1, pp. 62–67, 2021.

[41] I. O. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, A. Steiner, D. Keysers, J. Uszkoreit *et al.*, "Mlp-mixer: An all-mlp architecture for vision," *Advances in neural information processing systems*, vol. 34, pp. 24 261–24 272, 2021.

[42] C.-c. Jin and X. Chen, "An end-to-end framework combining time–frequency expert knowledge and modified transformer networks for vibration signal classification," *Expert Systems with Applications*, 2021.

[43] H. Liu, Y. Liu, Y. Wang, B. Liu, and X. Bao, "Eeg classification algorithm of motor imagery based on cnn-transformer fusion network," in *2022 IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*. IEEE, 2022.

[44] E. G. S. Nascimento, T. A. de Melo, and D. M. Moreira, "A transformer-based deep neural network with wavelet transform for forecasting wind speed and wind energy," *Energy*, vol. 278, p. 127678, 2023.

[45] M. I. Awan and J. Seokhee, "Surface texture classification based on transformer network," *Korean HCI Society Conference*, pp. 762–764, 2023.

[46] Y. Zou, P. Zou, Y. Zhao, K. Zhang, R. Zhang, and X. Wang, "Melons: generating melody with long-term structure using transformers and structure graph," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022.

[47] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang, "Informer: Beyond efficient transformer for long sequence time-series forecasting," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 12, 2021, pp. 11 106–11 115.

[48] G. Zerveas, S. Jayaraman, D. Patel, A. Bhamidipaty, and C. Eickhoff, "A transformer-based framework for multivariate time series representation learning," in *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, 2021, pp. 2114–2124.

[49] S. R. Khanal, E. V. Amorim, and V. Filipe, "Classification of car parts using deep neural network," in *CONTROLO 2020: Proceedings of the 14th APCA International Conference on Automatic Control and Soft Computing, July 1-3, 2020, Bragança, Portugal*. Springer, 2021, pp. 582–591.

[50] X. Lai, S. Zhang, N. Mao, J. Liu, and Q. Chen, "Kansei engineering for new energy vehicle exterior design: An internet big data mining approach," *Computers & Industrial Engineering*, vol. 165, p. 107913, 2022.

[51] F. Duvigneau, S. Liefold, M. Hoechstetter, J. L. Verhey, and U. Gabbert, "Analysis of simulated engine sounds using a psychoacoustic model," *Journal of Sound and Vibration*, vol. 366, pp. 544–555, 2016.

[52] M. Muender and C.-C. Carbon, "Howl, whirr, and whistle: The perception of electric powertrain noise and its importance for perceived quality in electrified vehicles," *Applied Acoustics*, vol. 185, p. 108412, 2022.

[53] P. Desmet, "Measuring emotion: Development and application of an instrument to measure emotional responses to products," *Funology 2: From Usability to Enjoyment*, pp. 391–404, 2018.

[54] X. Wu, M.-G. Yang, and Z. Su, "Pleasurable emotions of product design," *Intelligent Human Systems Integration (IHSI 2022): Integrating People and Intelligent Systems*, vol. 22, no. 22, 2022.

[55] K. Cha, "Affective scenarios in automotive design: a human-centred approach towards understanding of emotional experience," Ph.D. dissertation, Brunel University London, 2019.

[56] M. Grujicic, G. Arakere, V. Sellappan, J. Ziegert, and D. Schmueser, "Multi-disciplinary design optimization of a composite car door for structural performance, nvh, crashworthiness, durability and manufacturability," *Multidiscipline modeling in materials and structures*, 2009.

[57] Z. Wang, W. Liu, and M. Yang, "Data-driven multi-objective affective product design integrating three-dimensional form and color," *Neural Computing and Applications*, vol. 34, no. 18, pp. 15 835–15 861, 2022.

[58] S. J. Lederman and R. L. Klatzky, "Haptic perception: A tutorial," *Attention, Perception, & Psychophysics*, vol. 71, no. 7, pp. 1439–1459, 2009.

[59] C.-C. Carbon and M. Jakesch, "A model for haptic aesthetic processing and its implications for design," *Proceedings of the IEEE*, vol. 101, no. 9, pp. 2123–2133, 2012.

[60] M. Ajovalasit, R. Suriano, S. Ridolfi, R. Ciapponi, M. Levi, and S. Turri, "Human subjective response to aluminum coating surfaces," *Journal of Coatings Technology and Research*, vol. 16, pp. 791–805, 2019.

[61] F. Yin, R. Hayashi, R. Pongsathorn, and N. Masao, "Haptic velocity guidance system by accelerator pedal force control for enhancing eco-driving performance," in *Proceedings of the FISITA 2012 World Automotive Congress: Volume 12: Intelligent Transport System (ITS) & Internet of Vehicles*. Springer, 2013, pp. 37–49.

[62] D. Katzourakis, E. Velenis, E. Holweg, and R. Happee, "Haptic steering support for driving near the vehicle's handling limits; skid-pad case," *International Journal of Automotive Technology*, vol. 15, pp. 151–163, 2014.

[63] M. Lindemann, L. Nuy, K. Briele, and R. Schmitt, "Methodical data-driven integration of perceived quality into the product development process," *Procedia CIRP*, vol. 84, pp. 406–411, 2019.

[64] P. Ebel, F. Brokhausen, and A. Vogelsang, "The role and potentials of field user interaction data in the automotive ux development lifecycle: An industry perspective," in *12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 2020, pp. 141–150.

[65] S. Formentin, P. De Filippi, M. Corno, M. Tanelli, and S. M. Savaresi, "Data-driven design of braking control systems," *IEEE Transactions on Control Systems Technology*, vol. 21, no. 1, pp. 186–193, 2011.

[66] C. Sankavaram, B. Pattipati, A. Kodali, K. Pattipati, M. Azam, S. Kumar, and M. Pecht, "Model-based and data-driven prognosis of automotive and electronic systems," in *2009 IEEE international conference on automation science and engineering*. IEEE, 2009, pp. 96–101.

[67] X. Du and F. Zhu, "A new data-driven design methodology for mechanical systems with high dimensional design variables," *Advances in Engineering Software*, vol. 117, pp. 18–28, 2018.

[68] 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 City, United States: IEEE, 2016, pp. 393–395.

[69] Y. Yoo, I. Hwang, and S. Choi, "Consonance of vibrotactile chords," *IEEE transactions on haptics*, vol. 7, no. 1, pp. 3–13, 2013.

[70] K. Takahashi and J. Tan, "Deep visuo-tactile learning: Estimation of tactile properties from images," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 8951–8957.

[71] S. Okamoto, H. Nagano, and Y. Yamada, "Psychophysical dimensions of tactile perception of textures," *IEEE Transactions on Haptics*, vol. 6, no. 1, pp. 81–93, 2012.

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

[73] 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," *International Journal of Advanced Robotic Systems*, vol. 20, no. 4, p. 17298806231191947, 2023.

[74] D. Chen, D. Zhu, J. Liu, G. Chen, Y. Fang, and Y. Zhang, "Research on texture haptic reconstruction method based on informer model," in *Proceedings of the 2023 3rd International Conference on Robotics and Control Engineering*, 2023, pp. 161–165.

[75] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[76] S. Schiffman, "Introduction to multidimensional scaling: Theory, methods, and applications," 1981.

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

[78] A. Abdulali and S. Jeon, "Data-driven modeling of anisotropic haptic textures: Data segmentation and interpolation," in *International Conference on Human Haptic Sensing and Touch Enabled Computer Applications*. Springer, 2016, pp. 228–239.

[79] N. Landin, J. M. Romano, W. McMahan, and K. J. Kuchenbecker, "Dimensional reduction of high-frequency accelerations for haptic rendering," in *Haptics: Generating and Perceiving Tangible Sensations: International Conference, EuroHaptics 2010, Amsterdam, July 8-10, 2010. Proceedings*. Springer, 2010, pp. 79–86.

[80] H.-G. Kim, N. Moreau, and T. Sikora, *MPEG-7 audio and beyond: Audio content indexing and retrieval*. John Wiley & Sons, 2006.

[81] R. M. Haralick, K. Shanmugam, and I. H. Dinstein, "Textural features for image classification," *IEEE Transactions on systems, man, and cybernetics*, no. 6, pp. 610–621, 1973.

[82] M. Abadi and F. C. et al., "Tensorflow and keras: Open-source deep learning frameworks," https://www.tensorflow.org/ and https://keras.io/, 2015, accessed: 2024-03-13.

[83] P. Patil, Y. Wei, A. Rinaldo, and R. Tibshirani, "Uniform consistency of cross-validation estimators for high-dimensional ridge regression," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2021, pp. 3178–3186.

[84] V. W. Lumumba, D. Kiprotich, N. Makena, M. Kavita, and M. Mpaine, "Comparative analysis of cross-validation techniques: Loocv, k-folds cross-validation, and repeated k-folds cross-validation in machine learning models," *Am. J. Theor. Appl. Stat*, vol. 13, pp. 127–137, 2024.

[85] M. Stone, "Cross-validatory choice and assessment of statistical predictions," *Journal of the royal statistical society: Series B (Methodological)*, vol. 36, no. 2, pp. 111–133, 1974.

[86] G. T. Taye, H.-J. Hwang, and K. M. Lim, "Application of a convolutional neural network for predicting the occurrence of ventricular tachyarrhythmia using heart rate variability features," *Scientific reports*, vol. 10, no. 1, p. 6769, 2020.

[87] Z. Shao, J. Bao, J. Li, and H. Tang, "Haptic recognition of texture surfaces using semi-supervised feature learning based on sparse representation," *Cognitive Computation*, pp. 1656–1671, 2023.

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

[89] G. S. Giri, Y. Maddahi, and K. Zareinia, "An application-based review of haptics technology," *Robotics*, vol. 10, no. 1, p. 29, 2021.

[90] N. Heravi, W. Yuan, A. M. Okamura, and J. Bohg, "Learning an action-conditional model for haptic texture generation," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 11 088–11 095.

[91] T. H. Massie, "Initial haptic explorations with the phantom: Virtual touch through point interaction," Ph.D. dissertation, Massachusetts Institute of Technology, 1996.

[92] Fritz and K. E. Barner, "Stochastic models for haptic texture," in *Telemanipulator and Telepresence Technologies III*. SPIE, 1996.

[93] S. D. Cranstoun, H. C. Ombao, R. Von Sachs, Guo, and B. Litt, "Time-frequency spectral estimation of multichannel eeg using the auto-slex method," *IEEE transactions on Biomedical Engineering*, vol. 49, 2002.

[94] A. Abdulali, I. R. Atadjanov, and S. Jeon, "Visually guided acquisition of contact dynamics and case study in data-driven haptic texture modeling," *IEEE Transactions on Haptics*, vol. 13, no. 3, pp. 611–627, 2020.

[95] H. Culbertson, J. M. Romano, P. Castillo, M. Mintz, and K. J. Kuchenbecker, "Refined methods for creating realistic haptic virtual textures from tool-mediated contact acceleration data," in *2012 IEEE Haptics Symposium (HAPTICS)*. IEEE, 2012, pp. 385–391.

[96] W. Adi and S. Sulaiman, "Texture classification using wavelet extraction: An approach to haptic texture searching," in *2009 Innovative Technologies in Intelligent Systems and Industrial Applications*. IEEE, 2009, pp. 434–439.

[97] A. Jadon, A. Patil, and S. Jadon, "A comprehensive survey of regression based loss functions for time series forecasting," *arXiv preprint arXiv:2211.02989*, 2022.

[98] H. Hota, R. Handa, and A. K. Shrivas, "Time series data prediction using sliding window based rbf neural network," *International Journal of Computational Intelligence Research*, vol. 13, no. 5, pp. 1145–1156, 2017.

[99] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," in *International conference on machine learning*, 2013, pp. 1139–1147.

[100] W. McMahan and K. J. Kuchenbecker, "Dynamic modeling and control of voice-coil actuators for high-fidelity display of haptic vibrations," in *2014 IEEE Haptics Symposium (HAPTICS)*. IEEE, 2014, pp. 115–122.

[101] K. Tozuka and H. Igarashi, "A simplified texture modeling using a physical and perceptual rule-based approach," *IEEE Access*, 2024.

[102] D. Nie, J. Liu, and X. Sun, "Influence of surface tactile data quantity on material classification in unstructured environments," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–12, 2021.

[103] D. A. Norman, *Emotional design: Why we love (or hate) everyday things*. Civitas Books, 2004.

[104] P. Hekkert, "Design aesthetics: principles of pleasure in design," *Psychology Science*, vol. 48, no. 2, pp. 157–172, 2006.

[105] W. Kim, D. Park, Y. M. Kim, T. Ryu, and M. H. Yun, "Sound quality evaluation for vehicle door opening sound using psychoacoustic parameters," *Journal of Engineering Research*, vol. 6, no. 2, 2018.

[106] H. Brocke and S. Fayolle, "Haptics in product prototyping: potential and challenges," in *International Conference on Human Haptic Sensing and Touch Enabled Computer Applications*. Springer, 2009, pp. 230–239.

[107] B. Hoehn and et al., "Virtual prototyping in product development using immersive haptics," in *IEEE World Haptics Conference*. IEEE, 2009, pp. 475–480.

[108] J. Ma, J.-S. Kim, and K.-U. Kyung, "A hybrid haptic device for virtual car door interactions: Design and implementation," *IEEE Robotics and Automation Letters*, 2024.

[109] A. M. Okamura, "Haptics for robot-assisted minimally invasive surgery," *Current Opinion in Urology*, vol. 19, no. 1, pp. 102–107, 2009.

[110] T. Kim and et al., "Haptic feedback design for automotive rotary dial interfaces," *IEEE Transactions on Haptics*, vol. 12, no. 4, pp. 526–535, 2019.

[111] S. Shin, I. Lee, H. Lee, G. Han, K. Hong, S. Yim, J. Lee, Y. Park, B. K. Kang, D. H. Ryoo *et al.*, "Haptic simulation of refrigerator door," in *2012 IEEE Haptics Symposium (HAPTICS)*. IEEE, 2012, pp. 147–154.

[112] M. Nabeel, K. Aqeel, M. N. Ashraf, M. I. Awan, and M. Khurram, "Vibrotactile stimulation for 3d printed prosthetic hand," in *2016 2nd International Conference on Robotics and Artificial Intelligence (ICRAI)*. IEEE, 2016, pp. 202–207.

[113] K. Yoshida and H. Tanaka, "Development of a haptic virtual prototype for dishwasher handle simulation," in *IEEE World Haptics Conference*. IEEE, 2019, pp. 673–678.

[114] M. Kim and K.-U. Kyung, "Realistic haptic simulation of refrigerator handle interaction using hybrid actuator system," *IEEE Transactions on Industrial Electronics*, vol. 69, no. 2, pp. 1123–1133, 2022.

[115] J. Lee and S. Park, "Virtual prototyping of mechanical hand tools using task-specific haptic rendering," in *Proceedings of the ACM Symposium on User Interface Software and Technology*, 2022, pp. 211–222.

[116] 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," *IEEE Access*, pp. 1–1, 2025.

[117] M.-C. Bezat, V. Roussarie, T. Voinier, R. Kronland-Martinet, and S. Ystad, "Car door closure sounds: characterization of perceptual properties through analysis-synthesis approach," in *19th International Congress on Acoustics*, 2007, pp. CD–ROM.

[118] S. Yilmazer and Z. Bora, "Understanding the indoor soundscape in public transport spaces: A case study in akköprü metro station, ankara," *Building Acoustics*, vol. 24, no. 4, pp. 325–339, 2017.

[119] K. Priyadarshini, S. Chaudhuri, and S. Chaudhuri, "Perceptnet: Learning perceptual similarity of haptic textures in presence of unorderable triplets," in *Proc. IEEE World Haptics Conf.(WHC)*, 2019, pp. 163–168.

[120] J. Gwilliam and K. J. Kuchenbecker, "Kinesthetic vs. cutaneous force feedback for a stiffness discrimination task," in *2009 IEEE World Haptics Conference*. IEEE, 2009, pp. 97–102.

[121] K. E. MacLean, "Matching haptic vocabulary to human perception," in *Proceedings of EuroHaptics 2008*. Springer, 2008, pp. 2–6.

[122] B. T. Gleeson and W. R. Provancher, "Perceptual dimensions of tactile surface texture: A multidimensional scaling analysis," in *2010 IEEE Haptics Symposium*. IEEE, 2010, pp. 99–106.

# Appendix A

# List of Publications

### Journal Publications:

[1] **Awan, Mudassir Ibrahim,** and Seokhee Jeon. "Estimating Perceptual Attributes of Haptic Textures Using Visuo-Tactile Data." in IEEE Access, 2025, doi: 10.1109/ACCESS.2025.3581685.

[2] **Awan, Mudassir Ibrahim,** Ahsan Raza, Waseem Hassan, Ki-Uk Kyung, and Seokhee Jeon, "Quantifying Haptic Affection of Car Door through Data-Driven Analysis of Force Profile." in IEEE Access, 2025, doi: 10.1109/ACCESS.2025.3585067.

[3] **Awan, Mudassir Ibrahim,** Waseem Hassan, and Seokhee Jeon. "Fourier-enhanced Transformer Encoder Network for Efficient Haptic Texture Modeling/Rendering." **[Under Review]**

### Conference Publications:

[1] **Awan, Mudassir Ibrahim,** Myrah Naeem, and Seokhee Jeon. "Text-Driven Generative Framework for Multimodal Visual and Haptic Texture Synthesis." In 2025 IEEE World Haptics Conference (WHC). IEEE, 2025.

[2] **Awan, Mudassir Ibrahim**, et al. "Model-Mediated Teleoperation for Remote Haptic Texture Sharing: Initial Study of Online Texture Modeling and Rendering." 2023 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2023.

[3] **Awan, Mudassir Ibrahim,** Ahsan Raza, and Seokhee Jeon. "DroneHaptics: Encountered-Type Haptic Interface Using Dome-Shaped Drone for 3-DoF Force Feedback." In 2023 20th International Conference on Ubiquitous Robots (UR), pp. 195-200. IEEE, 2023.

[4] **Awan, Mudassir Ibrahim,** Waseem Hassan, and Seokhee Jeon. "Predicting perceptual haptic attributes of textured surface from tactile data based on deep cnn-lstm network." Proceedings of the 29th ACM Symposium on Virtual Reality Software and Technology. 2023.

[5] **Awan, Mudassir Ibrahim**, Seungchae Kim, Seokhee Jeon. (2024). "Haptic Feedback Chair for Simulating Heartbeat Sensations." 한국정보과학회 학술발표논문집, 전남.

[6] **Awan, Mudassir Ibrahim**, and Seokhee Jeon. "Surface texture classification based on transformer network." Proceedings of HCI Korea (2023): 762–764.

[7] Hashem, Mohammad Shadman, Ahsan Raza, **Mudassir Ibrahim Awan**, and Seokhee Jeon. "Pulsating Feedback: Render Human wrist Pulse via soft pneumatic actuator." 한국정보과학회 학술발표논문집 (2023): 1439-1441.

[8] **Awan, Mudassir Ibrahim**, and Seokhee Jeon. "Design and Evaluation of Lightweight Deep Learning Models for Synthesizing Haptic Surface Textures". 한국정보과학회 학술발표논문집, 제주 (2022)..

[9] Joolekha Bibi Joolee, Waseem Hassan, **Mudassir Ibrahim Awan**, Seokhee Jeon. "Haptic Texture Mapping on Real world 3D Object using Surface Texture and Haptic Model". 한국정보과학회 학술발표논문집, 제주 (2019).

## Nonreferred Demonstrations/Posters/Student Challenges:

[1] **Awan, Mudassir Ibrahim,** Ahsan Raza, Ji-Sung Kim, Jihyeong Ma, Jaehoon Chung, Ki-Uk Kyung and Seokhee Jeon. "Demonstration of a Hybrid Haptic Device and AI-Driven Car Door Profile Generation System." In IEEE World Haptics Conference (WHC), 2025. **(Demonstration)**

[2] **Awan, Mudassir Ibrahim,** and Seokhee Jeon. "HapTune: Demonstration of an Open-Source Visual Tool for Designing User-Defined Haptic Signals" In IEEE World Haptics Conference (WHC), 2025. **(Demonstration)**

[3] **Awan, Mudassir Ibrahim,** Jae-Ik Kim, Tae-Heon Yang and Seokhee Jeon. "Wearable Piezoelectric Tactile ring for Haptic Texture Rendering." In IEEE World Haptics Conference (WHC), 2025. **(Poster)**

[4] Raza, Ahsan, Mohammad Shadman Hashem, **Mudassir Ibrahim Awan,** and Seokhee Jeon. "Compact Multimodal Pneumatic Actuator Modules for Distributed Tactile Feedback." In IEEE World Haptics Conference (WHC), 2025. **(Poster)**

[5] **Awan, Mudassir Ibrahim,** and Seokhee Jeon. "Haptic Texture Rendering Using Transformer Encoder Network Integrated with Fourier Transforms." In Korea Haptics Conference, 2024. **(Demonstration)**

[6] **Awan, Mudassir Ibrahim,** Arsen Abdulali and Tatyana Ogay. "Haptic Cue: Providing Torque Response of Cue-ball Impact on the Wrist." In IEEE World Haptics Conference (WHC), 2019. **(Student Innovation Challenge – Virtual Reality)**

## Intellectual Properties:

[1] **Awan, Mudassir Ibrahim**, and Seokhee Jeon, "Dome shaped Haptic Drone for multi-directional Haptic interaction (돔 모양의 드론을 사용한 다차원 힘 촉감 제공 장치)." South Korean patent **(Domestic patent applied)**

[2] **Awan, Mudassir Ibrahim**, Tatyana Ogay, and Seokhee Jeon, "Realistic Tele-Sharing of Haptic Texture Using Model-Mediated and Adaptive Telepresence (촉각 질감의 사실적 원격 공유를 위한 모델기반 적응형 원격작업 시스템)." South Korean patent **(Domestic patent applied)**

[3] **Awan, Mudassir Ibrahim**, Tatyana Ogay, and Seokhee Jeon, "Apparatus and method for supporting tactile texture model-based remote control (촉각 질감 모델 기반 원격 제어 지원 장치 및 방법)." International patent **(US patent applied)**