Understanding touch through latent spaces: can images and haptic maps reflect human perception?

Antonio Luigi Stefani
University of Trento
antonioluigi.stefani@unitn.it

Sara Baldoni
University of Padova
sara.baldoni@unipd.it

Niccolò Bisagno
University of Trento
niccolo.bisagno@unitn.it

Federica Battisti
University of Padova
federica.battisti@unipd.it

Nicola Conci
University of Trento
nicola.conci@unitn.it

Francesco De Natale
University of Trento
francesco.denatale@unitn.it

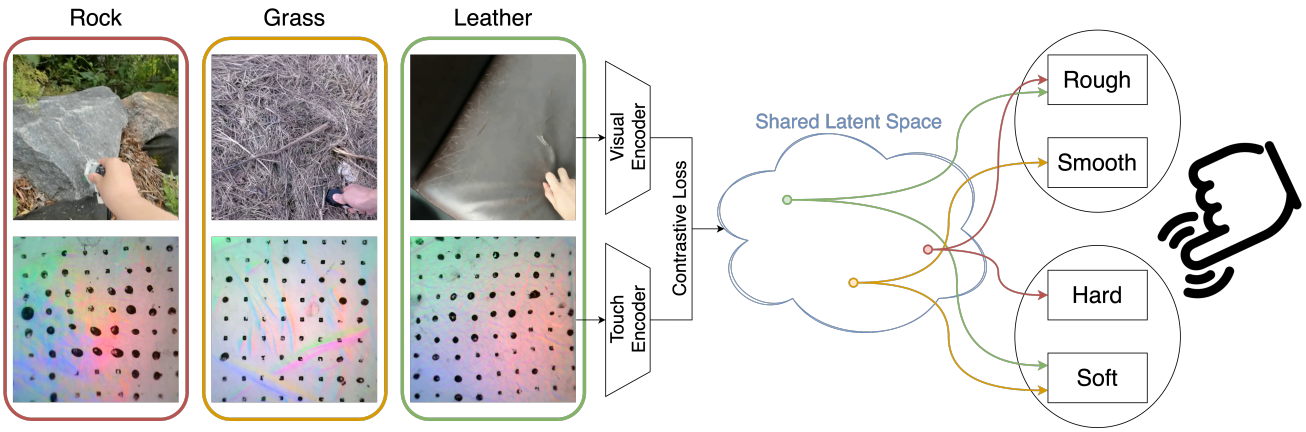


Figure 1. Overview of the proposed approach. On the left, paired RGB images and haptic maps from representative materials (e.g., rock, grass, leather) are processed by modality-specific encoders. On the right, the resulting embeddings are projected into a latent space and analyzed with respect to human perceptual dimensions such as rough/smooth and hard/soft. Our work investigates how visual and tactile modalities contribute to shaping a latent representation that aligns with human material perception.

Abstract

Extended Reality (XR) systems are increasingly incorporating multi-sensory stimuli to enhance realism and user immersion. Among these, the integration of tactile feedback plays a crucial role. Yet, the pipeline for acquiring, processing, and rendering haptic information—especially in synchrony with visual stimuli—remains largely unstandardized. A common strategy for capturing tactile data involves encoding it as haptic maps, essentially image-based representations of touch. However, the effectiveness of both visual and tactile modalities in modeling perceptual haptic properties is not yet fully understood.

In this study, we analyze the representational power of haptic maps and RGB images from the Touch and Go dataset using latent space analysis. Specifically, we investigate whether a neural network can structure the latent space in a way that reflects human perceptual attributes such as roughness, hardness, and colorfulness.

Our findings contribute to understanding whether haptic maps can serve as reliable proxies for tactile data and align with how humans perceive material properties, marking a step forward toward perceptually grounded haptic representations in XR environments.

1. Introduction

A characteristic aspect of Extended Reality (XR) systems, distinguishing them from traditional media, is the active interaction between users and virtual content. Users are no longer passive observers; they can move within the

scene, add or remove elements, and modify object features, thereby crafting personalized and immersive experiences.

Humans naturally engage with the world "multimodally", i.e., through multiple senses simultaneously. However, most current XR systems primarily rely on vision and hearing, neglecting other sensory modalities. To enhance realism and presence in virtual environments, it is essential to replicate the full range of human sensory experiences [1, 26]. In this context, touch plays a crucial role. The lack of tactile feedback—such as resistance when grabbing an object or variation in texture during surface interaction—significantly reduces perceived realism [5].

However, while visual and auditory integration are well-established, touch remains underexplored, lacking standard acquisition, processing, and rendering protocols [19]. This gap is further broadened by the incomplete understanding of how surface properties translate into human tactile perception. Despite the variability within each material class (e.g., a tree log versus a wooden table), humans are remarkably proficient at identifying materials and associating them with familiar categories and perceptual qualities such as roughness or softness. Moreover, humans can identify materials' characteristics (e.g., rough or smooth surfaces) and can relate new, unseen materials to familiar examples based on past experience and semantic information.

Material perception encompasses two interconnected tasks [9]: recognition, i.e., classifying a material into a semantic category (e.g., fabric or metal), and attribute assignment, i.e., associating physical qualities such as roughness, hardness, or warmth. While affective responses to touch (e.g., pleasantness) are highly subjective, physical properties have been studied more systematically and commonly include roughness (rough/smooth), hardness (hard/soft), and temperature (cold/warm) [15, 16].

Tactile data can be acquired using various sensors. In multimedia applications, they often take the form of mono-dimensional signals or 2D tactile maps [18]. These maps are produced by either inertial sensors—such as accelerometers and force sensors [12]—or by vision-based approaches [6, 28], which use cameras to capture the deformation of a membrane upon contact [14]. The output is an RGB image, commonly referred to as a tactile or haptic map, as illustrated in Fig. 2.

Although touch is central to material perception, vision also provides critical cues. As highlighted in [25], visual and tactile information contribute independently but complementarily to texture perception. Vision provides a significant contribution for identifying boundaries and coarse features, while touch excels at detecting fine-grained properties such as roughness. Moreover, visual input extends the perceptual space to include attributes such as glossiness and brightness [16]. This interplay underscores the importance of modeling both modalities to capture the full spectrum of material perception.

Although RGB images are widely used and understood to be rich in perceptual information, the extent to which they align with haptic features remains underexplored. Similarly, few studies have assessed whether haptic maps encode sufficient information to represent touch in a way that is meaningful for humans.

Our contribution. In this work, we explore whether a neural network can learn to represent both visual and tactile data in a latent space that aligns with human perception. We use contrastive learning—considering both supervised and self-supervised approaches to train models that embed RGB images and haptic maps from the Touch and Go dataset [27] into a shared latent space. Contrastive learning enables the model to distinguish similar (positive) from dissimilar (negative) pairs, promoting rich and semantically meaningful representations [23]. We then assess whether the resulting latent space reflects the perceptual dimensions identified in the tactile literature [15, 16].

This approach is a step toward perceptually grounded representations of tactile stimuli, which could eventually support automatic haptic rendering. Current haptic systems often rely on predefined parameters to simulate material properties; our work opens the possibility of learning these properties directly from perceptually aligned data. Our contributions can be summarized as:

- we investigate the relationship between haptic maps and human tactile perception;
- we extract both task-specific and task-agnostic feature representations for visual and touch using supervised and self-supervised contrastive learning approaches;
- we demonstrate that the learned latent space captures material properties in a way that aligns with established perceptual dimensions.

2. Related work

The alignment of haptic data and RGB images has been a recurring strategy since early haptic datasets [4, 21], enabling models to estimate tactile properties from visual input. In this context, haptic maps—image-based tactile representations—have gained relevance, as they allow leveraging computer vision techniques not suited to mono-dimensional signals. Despite their potential, XR applications have traditionally used simpler, one-dimensional tactile inputs. Recent efforts, however, have begun to explore the use of haptic maps for digitizing surfaces and objects in immersive scenarios.

Haptic maps for human-centered applications. Yang *et al.* [27] introduced one of the first datasets aligning RGB images and haptic maps with a focus on human interaction, moving beyond robot-centric setups involving small, gras-

pable objects. They proposed a contrastive learning framework to bring together features from RGB and tactile domains belonging to the same instance. This work forms the baseline for our experiments.

More recently, Dou *et al.* [7] introduced a dataset containing reconstructed 3D scenes—both indoor and outdoor—via Neural Radiance Fields (NeRFs), enabling the generation of novel-view haptic maps. Stefani *et al.* [20] further introduced a contact localization task to spatially anchor haptic feedback within 3D scenes.

Heravi *et al.* [11] adopted an action-conditioned approach using the HaTT dataset [4] to generate haptic feedback from tactile maps, connecting tactile data with embodied human interactions.

Although these works explore haptic maps from a human-centered perspective, it remains unclear whether such maps truly reflect human tactile perception. Moreover, it is not well understood whether neural networks trained on visual and haptic maps can approximate perceptual dimensions as experienced by humans.

Material perception in humans. To evaluate whether neural representations align with human perception, we draw on studies that systematically analyze human material perception.

Fleming *et al.* [9] conducted two comprehensive user studies to assess how people perceive material qualities. In the first, participants rated images from 10 material classes (e.g., glass, wood, metal) on nine perceptual attributes such as glossiness, roughness, and colorfulness using a six-point Likert scale. In the second, participants were given a list of 42 adjectives and were asked to associate them with six material categories, allowing for the construction of a material-perception space.

Awan *et al.* [2] expanded on this work by using physical samples of real materials (e.g., steel mesh, rubber, sandpaper). In the first phase of their study, participants selected the most relevant attributes (from a list of 25) for each material. This resulted in five primary perceptual dimensions: rough/smooth, flat/bumpy, sticky/slippery, hard/soft, and irritating/pleasant. In the second phase, participants rated each material on a continuous 0–100 scale along these dimensions.

In our study, we rely on these two sources [2, 9] to map the learned latent space to human perception. We focus on three dimensions: rough/smooth and hard/soft, which are common to both studies, and colorfulness, which is uniquely relevant to the visual domain. We exclude affective dimensions such as prettiness or pleasantness due to their high subjectivity.

In addition, we leverage the correlation matrix provided in the second experiment of [9], which quantifies how material categories relate in human perception. We compare these correlations with the distances among materials in our learned latent space to assess perceptual alignment.

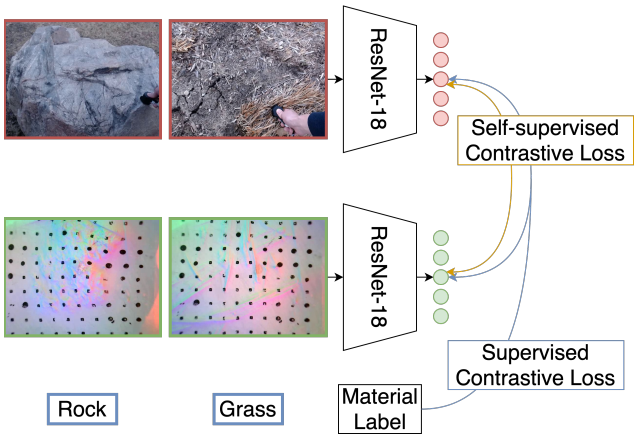


Figure 2. Self-supervised vs Supervised architectures.

3. Method

Our objective is to assess whether a neural network can learn visuo-tactile representations that reflect human perceptual dimensions—specifically, roughness, hardness, and colorfulness. To this end, we train both supervised and self-supervised contrastive learning models to map visual and tactile samples into a shared latent space. We then analyze this space to determine whether materials are organized in accordance with perceptual similarities.

Mapping visuo-tactile samples onto a shared latent space. We adopt two contrastive learning approaches—self-supervised and supervised—to learn a joint embedding space for visual and tactile inputs. These approaches aim to pull together paired samples from the two modalities and push apart unpaired ones.

Self-supervised methods [22, 27] are known for learning general-purpose features without requiring labels. In contrast, supervised contrastive learning [13] incorporates labeled data during training, often leading to more task-specific and less generalized feature representations.

Given a sample v_i from the visual dataset V and the corresponding sample t_i from the tactile dataset T, the objective is to represent paired samples (v_i, t_i) in a similar way, while pushing mismatched samples (v_i, t_j) apart in the representation space.

The network structure, illustrated in Fig. 2, consists of two ResNet18 encoders: one for the visual domain, $f_V(v_i)$, and one for the tactile domain, $f_T(t_i)$.

Following [13], we aim to optimize the following loss:

$$\mathcal{L} = \mathcal{L}_{V,T} + \mathcal{L}_{T,V}, \tag{1}$$

where each individual loss term encourages matching a visual sample v_i with a tactile sample and vice versa. Next, we separately define the different visual-to-tactile losses $\mathcal{L}_{V,T}$ for both cases. The tactile-to-visual $\mathcal{L}_{T,V}$ losses are not explicitly reported as they can be obtained by swapping the two domains.

For the self-supervised learning case, we employ a Contrastive Multiview Coding (CMC) model [22, 27], thus defining the following loss:

$$\mathcal{L}_{V,T} = -\log \frac{\exp(f_V(v_i) \cdot f_T(t_i)/\tau)}{\sum_{i \neq j} \exp(f_V(v_i) \cdot f_T(t_j)/\tau)}, \quad (2)$$

where $\tau = 0.07$ is a constant temperature parameter.

For the supervised version of our contrastive method, the visual-to-tactile loss is defined as:

$$\mathcal{L}_{V,T} = \sum_i -\frac{1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(f_V(v_i) \cdot f_T(t_i)/\tau)}{\sum_{i \neq j} \exp(f_V(v_i) \cdot f_T(t_j)/\tau)}, \quad (3)$$

where $P(i)$ represents the set of all positive sample indices in a batch, $|P(i)|$ denotes its cardinality, and τ is the same constant parameter.

Latent space analysis. After training, we extract features from the fifth ResNet layer, as commonly done in the contrastive learning literature [27], and apply t-distributed Stochastic Neighbor Embedding (t-SNE) [24] to project the high-dimensional embeddings into a 2D space. This visualization allows us to assess whether the learned representation clusters reflect perceptual dimensions relevant to human material understanding.

4. Experiments and discussion

Dataset. We conduct our tests on the Touch and Go dataset [27], the most comprehensive publicly available dataset providing paired visual and tactile samples with material annotations. Synchronized visual and tactile pairs are extracted from video frames, where the RGB images offer an egocentric perspective of the touch sample collection process, as shown by the samples in Fig. 2. The tactile samples are collected using a GelSight sensor [28], a vision-based tactile device in which a camera tracks the deformation of a curved elastomer gel, illuminated by multiple colored light sources and embedded with internal markers. The dataset comprises 13.9k detected touch samples of 3,971 individual object instances, categorized into 20 different classes. Among these, four classes are over-represented, with more than 1,000 samples each, while five classes are under-represented, with fewer than 400 samples each.

Implementation details. The self-supervised learning CMC model is trained in accordance with the experiments performed in [27] using a learning rate of 0.05 for 240 epochs. Stochastic Gradient Descent (SGD) is employed as optimizer, with a weight decay of 0.0001 and a momentum of 0.9. To fairly compare the self-supervised model, we trained our supervised learning model adopting the same hyperparameters. Both trainings are conducted with a batch size of 128 on a NVIDIA RTX 4090 GPU.

Experiments. To evaluate how the latent space is organized and whether it reflects human perception, we project the data into a shared latent space, which is analyzed through t-SNE visualizations. We evaluate the latent space representations obtained from visual and haptic inputs for different scopes, namely (A) material and perceptual attribute classification (B) alignment to human perception for hard/soft, rough/smooth, colorfulness features.

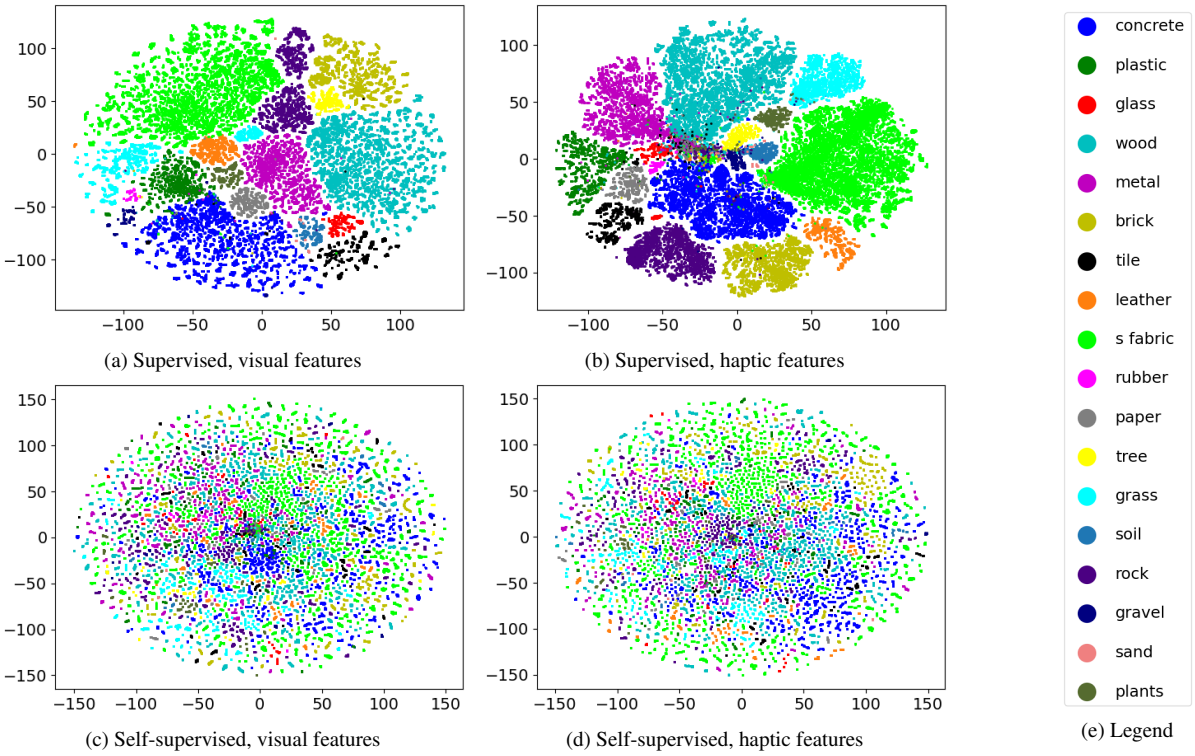
4.1. Latent space structure: supervised vs self-supervised

We first compare the ability of supervised and self-supervised contrastive learning frameworks to organize the latent space in a meaningful way. Our objective is to assess whether these models can encode visual and tactile features such that materials presenting similar perceptual attributes are placed nearby in the feature space.

To do so, we analyze the 2D t-SNE projections of the latent representations extracted from the fifth ResNet layer, as in [27]. As shown in Fig. 3, the self-supervised model fails to produce distinct clusters, whereas the supervised model—guided by material class labels—yields well-separated groupings across both visual and tactile modalities.

Visual vs. tactile representations. The latent space representations for the visual (Fig. 3a) and the tactile (Fig. 3b) domains present both commonalities and differences. Specifically, subsets of materials such as wood, metal, and glass, or grass and fabric, are close in both latent spaces, while others have different patterns. For instance, fabric and rock are close in the haptic latent space, while they are opposites in the visual latent space. Similarly, wood and rock are adjacent in the visual latent space, while they are distant in the touch space. A possible interpretation for this phenomenon is that the two encoders focus on different material features as occurs for the human perception channels [25]. Thus, as the touch domain is the primary focus of our study and following [27], we conduct our quantitative analysis on the material classification task using the tactile features.

Material classification. To quantify the difference between the supervised and the self-supervised approaches organization capabilities, we firstly performed material classification. The supervised model achieves an accuracy of 67.31%, while the self-supervised model obtains 54.7%.



(a) Supervised, visual features (b) Supervised, haptic features

(c) Self-supervised, visual features (d) Self-supervised, haptic features

(e) Legend

Figure 3. Supervised vision and touch vs self-supervised vision and touch latent space representations. Subfigures (a) and (c) (←) represent the material distribution in the latent space obtained from visual features; Subfigures (b) and (d) (→) represent the material distribution in the latent space obtained from tactile features, and Subfigure (e) contains the material-color legend.

This confirms the superior discriminative power of the supervised embeddings for material recognition. This phenomenon is in line with common learning tasks, where supervised approaches usually outperform unsupervised ones.

Perceptual attribute classification. We then evaluate whether the latent features encode perceptual properties beyond material identity. Specifically, we train binary classifiers to predict *rough/smooth* and *hard/soft* attributes. Unlike the previous experiment, this task tests whether the network organizes the latent space in alignment with perceptual judgments.

To this end, we consider two sources of labels:

- *Touch and Go labels*: dataset-provided annotations. *rough/smooth* labels are assigned per instance disregarding material classes, *hard/soft* labels are class-based and derived from sound cues [17].
- *Perceptual labels*: class-level labels based on human perception studies [2, 9]. For [9], we binarize the 6-point rating scale into rough/hard ([4–6]) and smooth/soft ([1–3]). For [2], a similar thresholding is applied to the 100-point scale (with [51–100] indicating the presence of a trait).

Touch and Go labels. We begin by training linear classifiers using the labels provided directly in the Touch and Go dataset. Classification results for both the rough/smooth and hard/soft dimensions are reported in Table 1. The supervised model consistently outperforms the self-supervised baseline, confirming its stronger ability to organize the latent space.

To better interpret these results, we visualize the latent space colored by the Touch and Go labels (Fig. 4). As seen in Fig. 4a and 4b, the supervised model produces a clear and consistent boundary between hard and soft materials in both modalities. The only exception is represented by the paper class which lies in the hard latent space portion, while being labeled as a soft material according to the dataset ground-truths. This phenomenon is class-specific: paper is typically associated with soft materials, however, the scanning process may lead to a misinterpretation of the hard attribute, as paper is often placed on hard surfaces such as tables or books.

In contrast, the self-supervised model fails to separate the two hardness categories clearly. A similar trend is observed for the rough/smooth dimension: while both models perform worse overall, the supervised tactile latent space (Fig. 4f) still exhibits two distinguishable clusters. This

Method	Accuracy	
	Roughness	Hardness
Supervised	**85.2**	**93.5**
Our self-supervised	82.0	78.9
Touch and Go [27] baseline	79.4	77.3

Table 1. Classification comparison between supervised and self-supervised methods on rough and hard attributes using the Touch and Go labels.

aligns with prior findings [25], where roughness perception is more strongly tied to touch than to vision—especially for fine-grained textures. The self-supervised tactile space (Fig. 4h) shows weaker structure but still encodes some perceptual signal.

These results also reflect the limitations of using instance-wise versus class-wise annotations. In particular, roughness—being highly variable within a material class—is poorly captured by class-level labels. Moreover, the dataset annotations do not always reflect human perception, which motivates our second experiment.

Perceptual labels. We train a binary linear classifier with the different sets of labels, using both supervised and self-supervised approaches. The results are shown in Table 2. Consistently, the supervised model outperforms the self-supervised one, especially for perceptual classification. This diverges from common trends in other domains [3, 10], where pre-training on self-supervised features often generalize better on downstream tasks thanks to the richer feature representation learned. In haptic, however, supervised learning appears to better capture semantically grounded perceptual cues.

Motivation for supervised model adoption. Our objective is not to directly compare the classification accuracy of the two learning frameworks, but to assess which model better organizes the latent space in alignment with human perception. The results presented above clearly demonstrate that the supervised model more accurately encodes both roughness and hardness as perceived by humans. It shows better boundary formation, more consistent clustering, and superior classification performance using both dataset-based and perception-based annotations.

Based on these findings, we adopt the supervised contrastive learning model for all subsequent experiments. This choice ensures that our latent space evaluations—especially those based on human perception—are grounded in the most semantically meaningful representation available.

Source	Method	Accuracy	
		Roughness	Hardness
[9]	Supervised	**99.75**	**99.73**
	Self-Supervised	89.90	94.06
[2]	Supervised	**99.69**	**99.86**
	Self-Supervised	93.56	97.72

Table 2. Classification comparison between supervised and self-supervised methods on rough and hard perceptual labels proposed respectively in [2, 9].

4.2. Human perception: a latent space evaluation

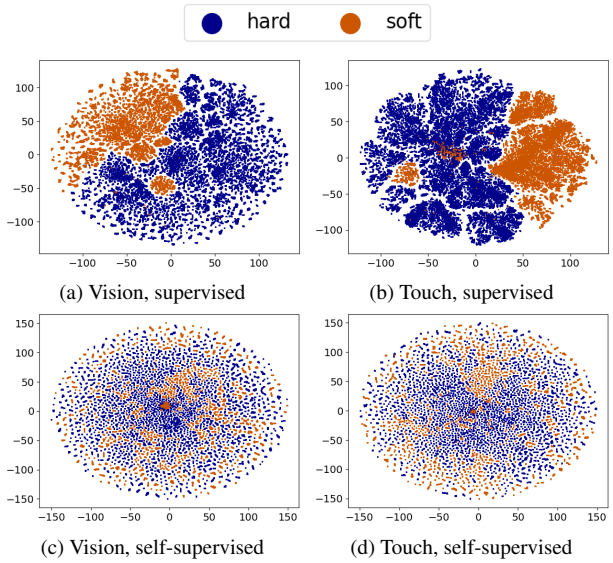
To assess whether the latent space representations align with human perceptual organization, as analyzed in [2, 9], we conduct three experiments by re-coloring the t-SNE projections according to perceptual classifications derived from these studies. In each visualization, darker colors (**P**) indicate material categories explicitly included in the original studies, while lighter tones (**P+**) represent materials that were not directly evaluated but were manually assigned to the closest perceptual class based on our interpretation.

Quantitative evaluations for these classifications are reported in Table 3, using the same linear classifier setup described in the previous section. This dual representation allows us to compare how well the latent space captures both directly validated perceptual labels and those inferred through extension.

Hard/Soft perceptual evaluation. Figures 5a and 5b present the hard/soft classification from [9], compared to our latent space projections in Fig. 4a and 4b. Here, the separation between hard and soft materials aligns almost perfectly, despite the fact that our models were trained using sound-derived labels from [17]. The main exception is leather, which is labeled as soft in Touch and Go and is located in the soft region of the latent space, but is perceived as hard in the perceptual study. This could be attributed to the specific type of leather used—for instance, flexible leather garments versus rigid leather upholstery.

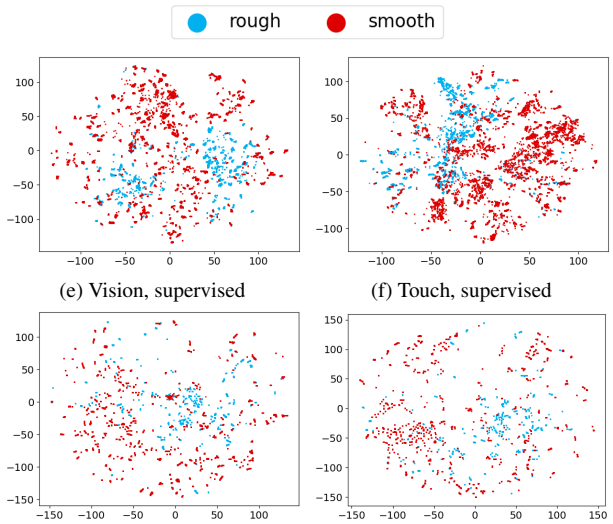
As to the perceptual analysis performed in [2], the subjective perception (Fig. 5d and 5c) is perfectly aligned to the latent space representation apart from the paper class, which is perceived as hard according to the perceptual study. Interestingly, although paper is labeled as *soft* in Touch and Go, it lies on the *hard* side of the boundary in the latent space, consistent with how it is perceived in the perceptual study. This further supports the claim that the learned representations, particularly in the supervised model, align well with human judgments.

These qualitative insights are supported by the quantitative results in Table 3, where the supervised classifier



(a) Vision, supervised (b) Touch, supervised

(c) Vision, self-supervised (d) Touch, self-supervised



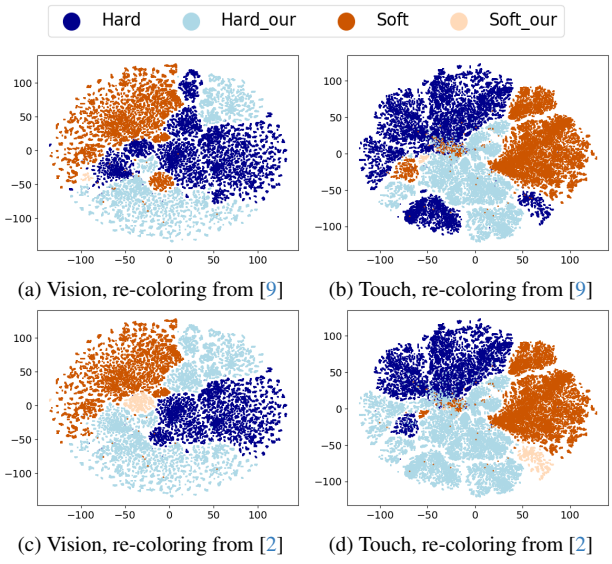
(e) Vision, supervised (f) Touch, supervised

(g) Vision, self-supervised (h) Touch, self-supervised

Figure 4. *Hard/soft* (↑) and *rough/smooth* (↓) samples distribution in the latent space colored according to the Touch and Go labels. Subfigures (a), (c), (e), and (g) (←) represent the visual feature distribution; Subfigures (b), (d), (f), and (h) (→) represent the tactile features distribution.

consistently achieves over 99% accuracy when training and testing on perceptual hard/soft labels (**P** and **P+**) from both studies. This confirms the strong alignment between the learned latent space and human perception of material compliance.

Rough/Smooth perceptual evaluation. Figures 6a and 6b show the latent space projections colored based on the roughness classification from [9]. These results reveal that the visual encoder does not form clearly separable



(a) Vision, re-coloring from [9] (b) Touch, re-coloring from [9]

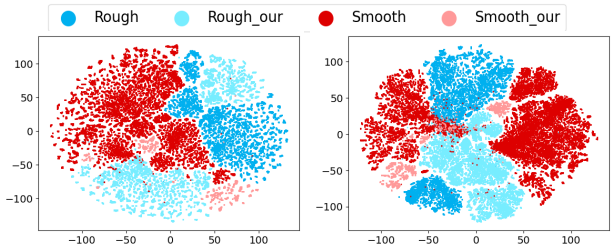
(c) Vision, re-coloring from [2] (d) Touch, re-coloring from [2]

Figure 5. Perception-based *hard/soft* classification. Subfigures (a) and (b) (↑) present the re-coloring based on [9], while subfigures (c) and (d) (↓) present the re-coloring based on [2].

Source	Train	Test	Accuracy	
			Roughness	Hardness
[9]	P	P	99.75	99.73
	P	P+	78.35	90.94
	P+	P	99.47	99.60
	P+	P+	99.54	99.57
[2]	P	P	99.69	99.86
	P	P+	76.71	87.77
	P+	P	99.51	99.83
	P+	P+	99.53	99.77

Table 3. Supervised classification accuracy across different combinations of training and testing perceptual labels. **P** refers to Perceptual classification labels proposed in the source work (darker colors), and **P+** refers to Perceptual classification including also our analysis.

clusters for rough and smooth materials. The tactile latent space, while better organized, still shows discrepancies with the perceptual labels. Specifically, plastic and paper are associated with the *rough* label while they are perceptually perceived as smooth according to [9]. Additionally, the rock class which is classified as *smooth* according to the Touch and Go dataset, is instead perceptually perceived as *rough*. However, this might strongly depend on the specific samples as natural rocks can be both rough and smooth, while the processed counterparts are usually smoother. In addition, it is interesting to note that the paper class was on average scored 3 out of 6 in the roughness-smooth



(a) Vision, re-coloring from [9] (b) Touch, re-coloring from [9]

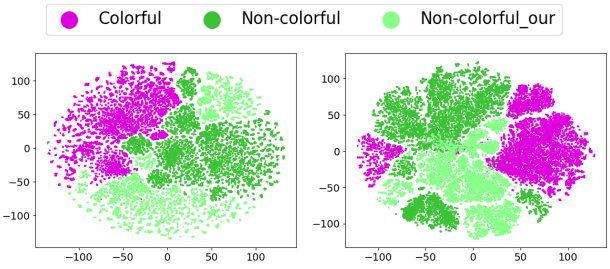
(c) Vision, re-coloring from [2] (d) Touch, re-coloring from [2]

Figure 6. Perception-based *rough/smooth* classification. Subfigures (a) and (b) (↑) present the re-coloring based on [9], while subfigures (c) and (d) (↓) present the re-coloring based on [2].

dimension in [9], thus it can be considered a challenging material to classify on the *rough/smooth* dimension based on human perception. This is further confirmed by the study performed in [2], where the paper material is associated to the *rough* class from a perceptual perspective (see Fig. 6c and 6d). As to the other materials, similar conclusions can be drawn. An exception is represented by fabric that is considered as *smooth* according to [9] and as *rough* according to [2]. This further confirms that the sample choice plays a relevant role in distinguishing between rough and smooth classes, and that this distinction should be sample-based rather than material-based to properly reflect human perception.

The roughness classification results in Table 3 reflect a more nuanced outcome. While accuracy remains high when training and testing within the same perceptual subset (e.g., **P** to **P**), performance drops significantly when generalizing to unseen samples (**P** to **P+**), especially for roughness. This highlights the subjective and sample-dependent nature of roughness perception and reinforces the need for instance-level labeling when modeling tactile texture.

Colorfulness evaluation. An additional dimension considered in [9] is colorfulness. The mapping between the perceptual classification and the latent space is provided in Fig. 7. It is interesting to note that while for the touch latent space there is not a clear separation, a boundary can be identified for the vision latent space. This result can be intuitively explained as color is a feature which can be encoded by the visual system while it is unperceivable through touch.



(a) Vision, re-coloring from [9] (b) Touch, re-coloring from [9]

Figure 7. Perception-based colorfulness classification.

Correlation of perceptual material classes. In the second experiment presented in [9], participants were asked to rate material classes using a variety of descriptive attributes, and the authors computed correlations among material categories. The study highlighted that there were significant correlations ($\rho > 0.35$) between the classes stone and wood, metal and stone, and glass and metal. It is possible to notice that these materials are adjacent in the visual latent space, while only glass and metal are close in the haptic latent space. This phenomenon can be motivated by the fact that, as it is well-established in the scientific literature [8], sight can be considered the humans' dominant sense. Since, during the second experiment, participants where not asked to touch any object but to associate adjectives to material classes, it is possible that the leading nature of sight has influenced the scoring of the participants. This might explain the closer relation of the vision latent space with the subjective evaluations with respect to the touch space.

5. Conclusions and Future Works

This paper explores whether a neural network can learn relationships between materials from visual and tactile inputs and structure a feature space aligned with human perception. We evaluate this using supervised and self-supervised contrastive learning. While the self-supervised model fails to organize the space meaningfully, the supervised one successfully clusters materials across both modalities. Through three perception-inspired experiments, we show that the extracted features resemble those used by humans, highlighting the importance of cues like roughness (haptic) and colorfulness (visual). However, our characterization of properties such as rough/smooth and hard/soft is currently class-level rather than instance-specific, limiting granularity. Moreover, although we relied on supervision, future work could investigate more generalizable self-supervised frameworks, such as Normalizing Flows, to structure the latent space in a perceptually meaningful way. These findings underscore the importance of multi-sensory integration in material perception and lay the groundwork for automatic, perceptually aligned stimulus generation in XR.

Acknowledgements

We acknowledge the support of the MUR PNRR project iNEST- Interconnected Nord-Est Innovation Ecosystem (ECS00000043) funded by the European Union under NextGenerationEU. In addition, this work was partially supported by the European Union under the Italian National Recovery and Resilience Plan (NRRP) Mission 4, Component 2, Investment 1.3, CUP C93C22005250001, partnership on "Telecommunications of the Future" (PE00000001 - program "RESTART").

References

[1] Sepehr Alizadehsalehi, Ahmad Hadavi, and Joseph Chuenhuei Huang. From BIM to extended reality in AEC industry. *Automation in construction*, 116:103254, 2020. 2

[2] Mudassir Ibrahim Awan, Waseem Hassan, and Seokhee 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*, New York, NY, USA, 2023. Association for Computing Machinery. 3, 5, 6, 7, 8

[3] Mathilde Caron, Piotr Bojanowski, Julien Mairal, and Armand Joulin. Unsupervised pre-training of image features on non-curated data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2959–2968, 2019. 6

[4] Heather Culbertson, Juan José López Delgado, and Katherine J Kuchenbecker. One hundred data-driven haptic texture models and open-source methods for rendering on 3d objects. In *2014 IEEE haptics symposium (HAPTICS)*, pages 319–325. IEEE, 2014. 2, 3

[5] Aleph Campos Da Silveira and Celso Alberto Saibel Santos. Ongoing challenges of evaluating mulsemedia QoE. In *Workshop on Multisensory Experiences (SensoryX)*. SBC, 2022. 2

[6] Won Kyung Do and Monroe Kennedy. Densetact: Optical tactile sensor for dense shape reconstruction. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 6188–6194. IEEE, 2022. 2

[7] Yiming Dou, Fengyu Yang, Yi Liu, Antonio Loquercio, and Andrew Owens. Tactile-augmented radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26529–26539, 2024. 3

[8] Jamie Enoch, Lee Jones, and Leanne McDonald. Thinking about sight as a sense. *Optometry in Practice*, 21(3):2–9, 2020. 8

[9] Roland W. Fleming, Christiane Wiebel, and Karl Gegenfurtner. Perceptual qualities and material classes. *Journal of Vision*, 13(8):9–9, 2013. 2, 3, 5, 6, 7, 8

[10] Jiawei Ge, Shange Tang, Jianqing Fan, and Chi Jin. On the provable advantage of unsupervised pretraining. *arXiv preprint arXiv:2303.01566*, 2023. 6

[11] Negin Heravi, Heather Culbertson, Allison M Okamura, and Jeannette Bohg. Development and evaluation of a learning-based model for real-time haptic texture rendering. *IEEE Transactions on Haptics*, 2024. 3

[12] Zhanat Kappassov, Juan-Antonio Corrales, and Véronique Perdereau. Tactile sensing in dexterous robot hands — Review. *Robotics and Autonomous Systems*, 74:195–220, 2015. 2

[13] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020. 3

[14] Mike Lambeta, Po-Wei Chou, Stephen Tian, Brian Yang, Benjamin Maloon, Victoria Rose Most, Dave Stroud, Raymond Santos, Ahmad Byagowi, Gregg Kammerer, et al. Digit: A novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation. *IEEE Robotics and Automation Letters*, 5(3):3838–3845, 2020. 2

[15] Justin D. Lieber and Sliman J. Bensmaia. The neural basis of tactile texture perception. *Current Opinion in Neurobiology*, 76:102621, 2022. 2

[16] Shogo Okamoto, Hikaru Nagano, and Yoji Yamada. Psychophysical Dimensions of Tactile Perception of Textures. *IEEE Transactions on Haptics*, 6(1):81–93, 2013. 2

[17] Andrew Owens, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H Adelson, and William T Freeman. Visually indicated sounds. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2405–2413, 2016. 5, 6

[18] Aaron Raymond See, Jose Antonio G Choco, and Kohila Chandramohan. Touch, texture and haptic feedback: a review on how we feel the world around us. *Applied Sciences*, 12(9):4686, 2022. 2

[19] Antonio Luigi Stefani, Niccolò Bisagno, Andrea Rosani, Nicola Conci, and Francesco De Natale. Signal Processing for Haptic Surface Modeling: a Review. *arXiv preprint arXiv:2409.20142*, 2024. 2

[20] Antonio Luigi Stefani, Niccolò Bisagno, Nicola Conci, and Francesco De Natale. Splattouch: Explicit 3d representation binding vision and touch. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 118–127, 2025. 3

[21] Matti Strese, Jun-Yong Lee, Clemens Schuwerk, Qingfu Han, Hyoung-Gook Kim, and Eckehard Steinbach. A haptic texture database for tool-mediated texture recognition and classification. In *2014 IEEE International Symposium on Haptic, Audio and Visual Environments and Games (HAVE) Proceedings*, pages 118–123. IEEE, 2014. 2

[22] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 776–794. Springer, 2020. 3, 4

[23] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? *Advances in neural information processing systems*, 33:6827–6839, 2020. 2

[24] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of machine learning research*, 9 (11), 2008. 4

[25] T. Aisling Whitaker, Cristina Simões-Franklin, and Fiona N. Newell. Vision and touch: Independent or integrated systems for the perception of texture? *Brain Research*, 1242:59–72, 2008. Multisensory Integration. 2, 4, 6

[26] Nannan Xi, Juan Chen, Filipe Gama, Marc Riar, and Juho Hamari. The challenges of entering the metaverse: An experiment on the effect of extended reality on workload. *Information Systems Frontiers*, 25(2):659–680, 2023. 2

[27] Fengyu Yang, Chenyang Ma, Jiacheng Zhang, Jing Zhu, Wenzhen Yuan, and Andrew Owens. Touch and go: Learning from human-collected vision and touch. *arXiv preprint arXiv:2211.12498*, 2022. 2, 3, 4, 6

[28] Wenzhen Yuan, Siyuan Dong, and Edward H Adelson. Gelsight: High-resolution robot tactile sensors for estimating geometry and force. *Sensors*, 17(12):2762, 2017. 2, 4