Material Classification using Haptic and Vision Data

Karan Baijal
College of Engineering
Cornell University
United States
kb553@cornell.edu

1 Introduction

Material classification plays a vital role in robotic manipulation and task planning, enabling robots to interact with objects safely and effectively. Accurate material recognition is critical for differentiating between visually similar objects that may have vastly different functional requirements or to identify the appropriate action to take for object manipulation. For instance, during task planning, when given a natural language instruction "Place the bottle in the microwave", the robot must distinguish between an aluminum bottle, which is unsuitable for microwaving, and a plastic one, which is appropriate. Similarly, in manipulation tasks, recognizing material properties—such as differentiating between a ceramic and a plastic plate—allows robots to adjust their grip force and handling strategy, minimizing the risk of damage or failure.

Current material recognition systems in robotics predominantly rely on visual data. While vision-based methods perform reasonably well in structured and well-lit environments, they often struggle in real-world settings, such as low-light conditions, cluttered environments, or when dealing with visually similar objects, where visual cues may be insufficient. For robots to effectively manipulate objects in complex environments, such as household settings, integrating haptic sensing — a combination of touch and kinesthesia — is essential. For example, in a task like "Grab the apple for cooking," haptic sensing can discern between a real apple and an artificial one by detecting texture and firmness, overcoming the limitations of visual methods.

Despite the potential of haptic sensing, prior research in material recognition using haptics is limited. Much of the existing work relies on expensive sensors or vision-centric approaches, which perform well only in controlled environments and struggle to generalize to new objects in unstructured settings. Additionally, they usually employ traditional machine learning tools for material recognition rather than the latest deep learning models. In this work, we aim to solve this problem through exploring the following hypotheses:

- Integrating vision and haptic data enhances material recognition accuracy compared to using either modality alone.
- Leveraging haptic data enables the development of multi-encoder machine learning models that outperform purely end-to-end approaches by incorporating the physics of materials and sensor behavior.
- Deep-learning models enable one-shot material recognition of in-the-wild household objects.

To validate these hypotheses, we utilize the MIDAS-Set, a crowdsourced dataset collected with the Multimodal In-the-wild Haptic Data Acquisition System (MIDAS). This dataset combines vision and haptic sensor data from household objects, offering a challenging and diverse benchmark for advancing material recognition research.

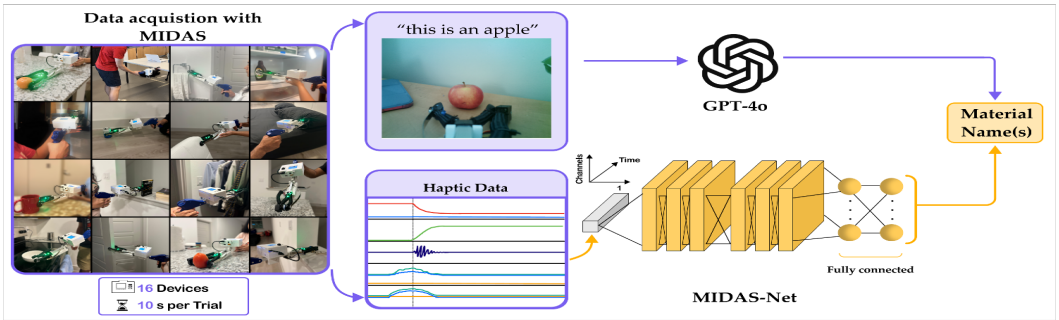


Figure 1: Overarching Material Classification Framework

2 Related Work

Previous research has utilized haptic modalities to infer object properties [1], recognize textures [2, 3], and classify materials [4, 5, 6] and objects [7, 8]. These studies demonstrate the potential of haptic data in tasks where tactile feedback provides critical information that vision alone cannot achieve.

There has also been previous work in using vision for material classification. Drehwald et al. [9] trained CLIP on a combined dataset of synthetic and natural images for one-shot material recognition. Similarly, Bell et al. [10] utilized a convolutional neural network (CNN) trained on a large-scale internet-sourced image dataset for material classification. However, these methods exclusively rely on visual information, making them susceptible to challenges such as visually similar objects, occlusions, or poor lighting conditions. Additionally, they do not incorporate haptic information, which could significantly enhance material recognition, nor do they address the complexities of working with crowdsourced images of household objects.

Some prior methods have fused haptic information with vision [11, 12, 13]. Luo et al. [14] provide an in-depth review of haptic perception towards object property recognition. Most existing work relies on small, curated haptic datasets [1], limiting the generalizability of these models to real-world, unstructured environments. In contrast, our work leverages the MIDAS-Set, a crowdsourced, in-the-wild dataset, enabling more robust and scalable predictions of material properties across a diverse range of household objects.

Data-driven methods in prior work have used classical machine learning methods, such as Hidden Markov Models (HMMs), Support Vector Machines, and [5, 15, 16] to classify materials using time-series haptic data, however these methods struggle to scale effectively with dataset size and diversity. Recent work has used large models to classify haptic signals [17, 18]. In this work, we explore the use of temporal convolutional network-based architecture and transformers to perform material recognition from in-the-wild haptic data.

3 Problem Formulation

The MIDAS-Set dataset, used in this work, comprises 15,000 samples of raw haptic sensor data collected from the MIDAS device. Each sample consists of time-series data captured by multiple sensors: a force sensor, a contact microphone, a passive temperature sensor, an active temperature sensor, and two inertial measurement units (IMUs), as shown in Figure 2. In addition to these raw sensory inputs, we introduce a derived feature, Impedance, which is computed as the ratio of the force sensor reading to the angular velocity of the arm during a grasp. This feature captures dynamic interactions between the robotic arm and the object being manipulated.

Thus, the input to our haptic model is represented as a tensor of shape $15,000 \times 5 \times 420$, corresponding to 15,000 samples, 5 features (Force, Impedance, Contact Microphone, Active Temperature, and Passive Temperature), and 500 timesteps (sensors sampled at 50hz over a 10-second interval). For

the vision model, the inputs were images of the objects corresponding to the haptic data, preprocessed using Grounded SAM [19][20] to remove distractor elements (described in Methods).

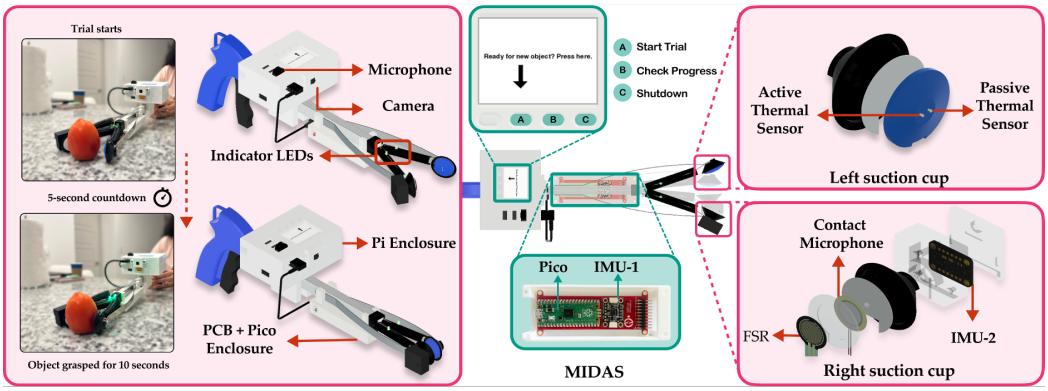


Figure 2: Device Overview: MIDAS device sensors used for data collection.

The output of the model is a material label that categorizes a given household object into one of the following predefined material classes: ['aluminium', 'brass', 'cardboard', 'fabric', 'foam', 'glass', 'granite','hard_plastic', 'paper', 'porcelain', 'rubber', 'soft_plastic', 'steel', 'vegetable', 'wood']. This closed-vocabulary set aligns with the annotations provided in the dataset. For our model learning, we assume MIDAS-Set contains accurate labels and vision and haptic data.

To train and evaluate the model, we use an 80:10:10 split of the dataset for training, validation, and testing, respectively. Model performance on the test set is assessed using the following evaluation metrics:

- Class Accuracy: The proportion of correctly classified samples, averaged across all material classes. This metric is defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

- Normalized Matthews Correlation Coefficient (NMCC) Score: The MCC evaluates the overall quality of predictions by considering all four components of the confusion matrix. Unlike metrics such as F1 score, which ignore true negatives, the MCC provides a more balanced view of classifier performance, especially for imbalanced datasets. The NMCC is a normalized version of the MCC, rescaling its values from the original range of $[-1, 1]$ to $[0, 1]$. It is defined as:

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{2}$$

$$\text{NMCC} = \frac{1 + MCC}{2} \tag{3}$$

In the above equations, TP is True Positive, TN is True Negative, FP is False Positive, and FN is False Negative. Given the diverse and imbalanced nature of the dataset, we expect the MCC score to serve as a more representative metric of model performance compared to class accuracy.

4 Method

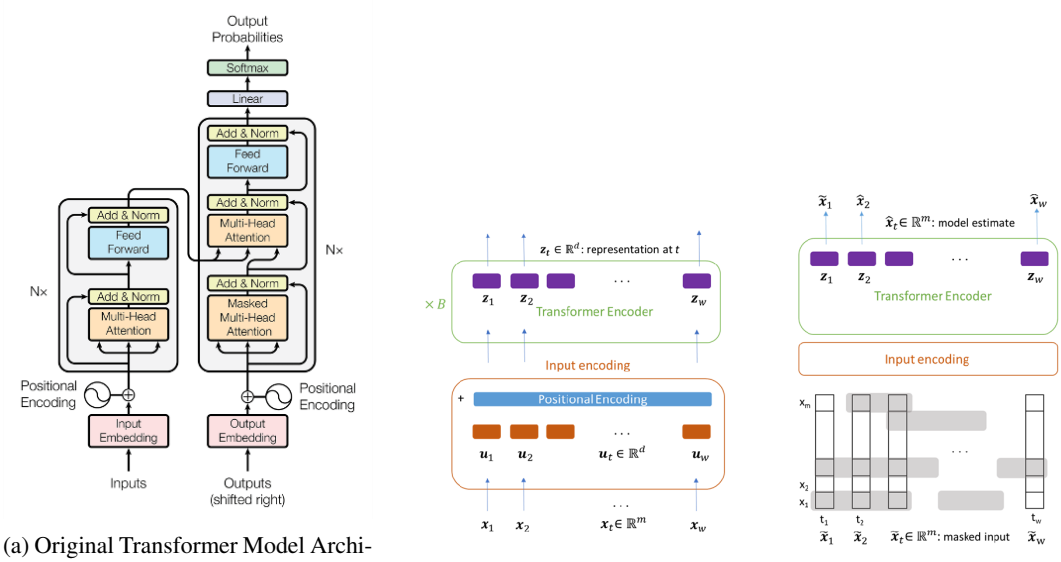
This work explores advanced deep learning architectures for material recognition by leveraging both haptic and vision data. For haptic learning, we investigate the use of transformers and convolutional neural networks (CNNs) in both multi-encoder and end-to-end configurations. For vision learning, we utilize two state-of-the-art Vision-Language models: OpenAI CLIP and GPT-4o.

Across all models (except GPT-4o), we apply weighted categorical cross-entropy loss with label smoothing ($\epsilon = 0.1$) to mitigate the effects of label noise. To address the dataset's long-tailed class distribution, we use class weights derived from the square root of the inverse class population.

4.1 Haptic Learning

4.1.1 End-to-End Model

Multivariate Time-Series Transformer: We adopt a modified version of the Multivariate Time-Series Transformer developed by Zervas et al. [21]. First, time series inputs (x_t) are encoded into a higher-dimensional space (u_t) using a combination of input encodings and positional encodings to retain temporal and sequential information. These embeddings are passed into a Transformer Encoder, which models temporal dependencies and relationships across features using self-attention mechanisms (Figure 3(a)), producing contextual representations (z_t) for each time step. The final representations ($\tilde{x_t}$) are obtained after decoding and used for classification, as shown in Figure 3(b). This design ensures effective temporal modeling and feature interaction across modalities.



(a) Original Transformer Model Architecture [22]

(b) Multivariate Time-series Transformer [21]

Figure 3: Transformer Architecture

Temporal CNN: We use a modified version of Inception Time architecture [23], shown in Figure 4. InceptionTime applies 1-D convolutions with variable-length filters to time-series data. The model includes six Inception blocks, each with 512 filters. Its hierarchical design enables multi-scale feature extraction, making it particularly suitable for time-series data.

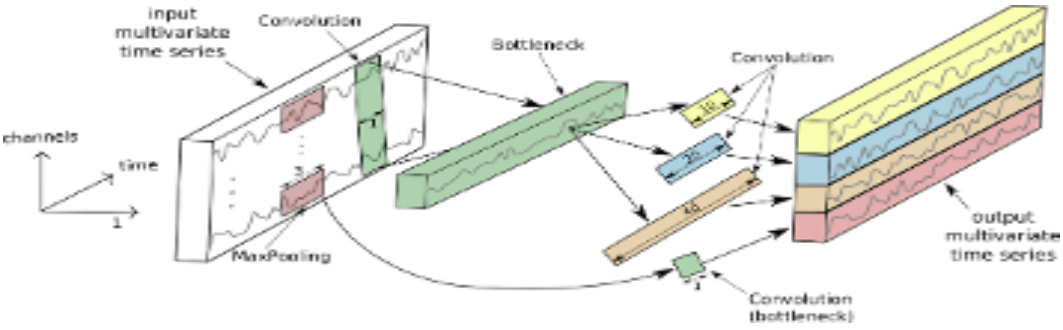


Figure 4: InceptionTime Model Architecture [23]

4.1.2 Multi-encoder Network

Multi-Encoder Haptic Transformer: We extend the end-to-end transformer architecture by introducing modality-specific encoders to better handle distinct sensor modalities. We develop the attention layers such that self-attention is applied to each modality to capture intra-modality dependencies and cross-attention layers are applied between conditionally dependent features. For example, Force sensor, IMU-2, and contact microphone data are grouped to measure object hardness, and cross-attention is applied to these features. Similarly, active and passive temperature sensors are grouped to measure thermal conductivity, and cross-attention is applied to these features, as shown in Figure 5.

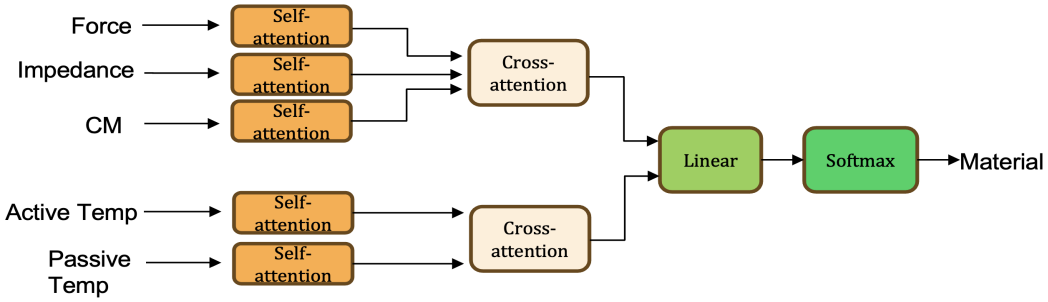


Figure 5: Haptic Transformer Architecture

Multi-encoder Haptic Temporal CNN (TCNN): We design a multi-encoder architecture to model the relationships between different sensor modalities, similar to the haptic transformer. This workflow includes separate TCNN blocks for related sensor groups, e.g., force impedance and contact microphone are processed together, while active and passive temperature sensors are processed as a separate block. Outputs from these modality-specific blocks are combined and passed through another TCNN block to generate the final material label, as shown in Figure 6.

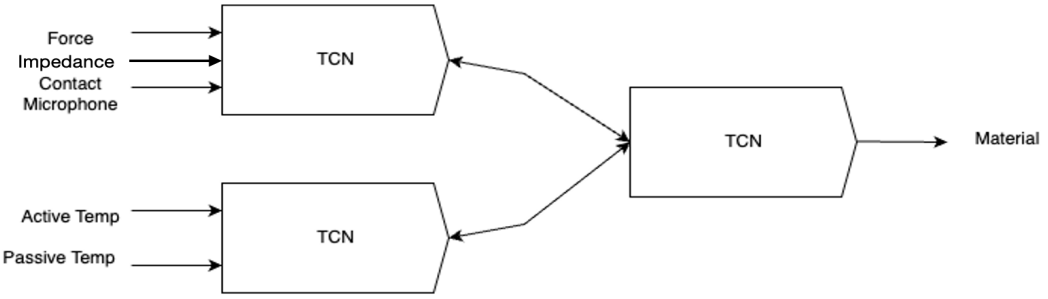


Figure 6: Haptic CNN Architecture

4.2 Vision Learning

For vision-based material classification, we preprocess images using Grounded SAM to remove distractor items and MIDAS grippers from the input images. The processed images are evaluated using two Vision-Language models: OpenAI CLIP and GPT-4o.

CLIP: We fine-tune the last six layers of the CLIP model and append a two-layer multilayer perceptron (MLP) followed by a final fully connected layer for material classification. Fine-tuning focuses on leveraging pre-trained visual and textual embeddings to adapt the model to the MIDAS-Set dataset.

GPT-4o: We employ a two-step prompting approach for better material classification. We initially feed in an image to identify the object in question. Then, we further prompt GPT with the determined

object and input image to use its common-sense reasoning and textual understanding of the image for material classification.

5 Experiments

5.1 Experimental Setup

For data pre-processing, we time-synchronize all sensor data to within 2 ms across modalities and smooth the active and passive temperature signals using a moving average filter and a zero-phase digital filter.

This processed data is provided as input to the haptic models. Each model was run three times with different test sets for a more robust evaluation. The average performance and standard deviations across these runs are reported for each model.

5.2 Quantitative Results

The quantitative results of the experiments are summarized in Tables 1 and 3.

Model	Accuracy	NMCC
Transformer	0.43 ± 0.05	0.60 ± 0.02
Transformer (Separate Encoders)	0.47 ± 0.04	0.61 ± 0.03
TCNN	0.61 ± 0.005	0.77 ± 0.005
TCNN (Separate Encoders)	0.43 ± 0.015	0.67 ± 0.02

Table 1: Haptic Model Evaluation Metrics

With both models, end-to-end models performed better than multi-encoder networks. Additionally, TCNN outperforms the Transformer model significantly.

Model	Accuracy	NMCC
CLIP	0.48 ± 0.007	0.68 ± 0.01
GPT-4o	0.51 ± 0.05	0.73 ± 0.02

Table 2: Vision Model Evaluation Metrics

For material recognition using purely vision, GPT-4o outperforms CLIP across all metrics.

Model	Accuracy	NMCC
Haptics only (TCNN)	0.61 ± 0.005	0.77 ± 0.005
Vision Only (GPT-4o)	0.51 ± 0.05	0.73 ± 0.02
Haptic+Vision	0.82 ± 0.04	0.91 ± 0.03

Table 3: Material Recognition Evaluation Metrics

Combining haptic and vision data significantly improves material classification performance, achieving an accuracy of 82% and an NMCC of 0.91. The haptic-only model also performs well, outperforming the vision-only model.

5.3 Qualitative Results

The results from Table 1 demonstrate that end-to-end architectures, such as the TCNN, outperform separate encoder-based architectures for haptic data processing. This indicates that the sensor modalities (e.g., force, impedance, and thermal signals) are interdependent, and an end-to-end design is more effective in capturing their relationships. The transformer model did not do as strongly as expected and overfit quickly to the training data. This suggests that more data may be needed for better Transformer performance. In table 2, we notice that GPT outperforms CLIP. This may

signal to the inherent importance of semantic and contextual information in vision-based material classification that GPT provides.

From Table 3, it is evident that the combination of haptic and vision data leads to the best material recognition performance. While the haptic model alone performs well, integrating vision data enhances the model's ability to differentiate between objects with subtle material differences. Lastly, the haptic + vision model demonstrates strong one-shot material recognition capabilities. The test set includes unseen, in-the-wild household objects, and the model's ability to generalize to these scenarios underscores the robustness of the proposed approach.

6 Conclusion

In this work, we presented a novel approach to material classification by integrating haptic and vision data, addressing challenges in recognizing materials in real-world household environments. Our experiments demonstrate that end-to-end models outperform multi-encoder designs, highlighting the importance of capturing feature dependencies across modalities. By combining vision and haptic data, we achieve superior classification performance, leveraging the strengths of both modalities for material recognition. We further enable one-shot material recognition using our proposed architecture, showcasing the model's ability to generalize from limited examples. We provide the following novel contributions:

- **Feature Dependency Analysis in Haptic Data Learning:** We demonstrate that end-to-end models outperform separate encoder architectures for haptics by effectively capturing inter-dependencies between modalities.
- **Integration of Vision and Haptics for Enhanced Material Classification:** By combining haptic and vision data, we achieve improved performance, leveraging the complementary strengths of both modalities.
- **One-Shot Material Recognition Capability:** Our model enables one-shot recognition of materials, allowing for accurate identification with minimal labeled data.

Moving forward, we are currently deploying the model on a robotic platform. We are replacing the MIDAS gripper with a linear servo motor to control the grip speed of our device and replace with the Franka Emika Panda robot arm end effector. This deployment will allow us to evaluate whether haptic data collected manually remains in distribution during real-world robotic tasks, ensuring reliable material classification under practical conditions.

Despite promising results, this work has limitations. The haptic data used in this study was collected with inexpensive sensors in a crowdsourced setup, introducing noise that may have impacted model performance. Data collected from higher-quality sensors in a controlled laboratory environment could further improve accuracy and generalization. While this work primarily focused on material classification, future research could explore integrating material recognition into control and task planning frameworks, enabling robots to make informed decisions based on material properties - similar to problems mentioned in the introduction.

In summary, our work highlights the potential of multimodal approaches for robust material recognition and lays the foundation for their integration into real-world robotic systems.

References

[1] V. Chu, I. McMahon, L. Riano, C. G. McDonald, Q. He, J. M. Perez-Tejada, M. Arrigo, N. Fitter, J. C. Nappo, T. Darrell, et al. Using robotic exploratory procedures to learn the meaning of haptic adjectives. In *2013 IEEE International Conference on Robotics and Automation*, pages 3048–3055. IEEE, 2013.

[2] J. Sinapov, V. Sukhoy, R. Sahai, and A. Stoytchev. Vibrotactile recognition and categorization of surfaces by a humanoid robot. *IEEE Transactions on Robotics*, 27(3):488–497, 2011.

[3] T. Taunyazov, Y. Chua, R. Gao, H. Soh, and Y. Wu. Fast texture classification using tactile neural coding and spiking neural network. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9890–9895. IEEE, 2020.

[4] T. Bhattacharjee, J. Wade, and C. C. Kemp. Material recognition from heat transfer given varying initial conditions and short-duration contact. In *Robotics: Science and Systems*, volume 2015, 2015.

[5] T. Bhattacharjee, H. M. Clever, J. Wade, and C. C. Kemp. Multimodal tactile perception of objects in a real home. *IEEE Robotics and Automation Letters*, 3(3):2523–2530, 2018.

[6] Z. Erickson, S. Chernova, and C. C. Kemp. Semi-supervised haptic material recognition for robots using generative adversarial networks. In *Conference on Robot Learning*, pages 157–166. PMLR, 2017.

[7] L. E. R. Bonner, D. D. Buhl, K. Kristensen, and N. Navarro-Guerrero. Au dataset for visuo-haptic object recognition for robots. *arXiv preprint arXiv:2112.13761*, 2021.

[8] N. Gorges, S. Escaida Navarro, D. Göger, and H. Wörn. Haptic object recognition using passive joints and haptic key features. In *2010 IEEE International Conference on Robotics and Automation*, pages 2349–2355, 2010. doi:10.1109/ROBOT.2010.5509553.

[9] M. Drehwald, K. Krösl, B. Holzschuh, and K. Bühler. One-shot recognition of any material anywhere using contrastive learning with physics-based rendering. *ACM Transactions on Graphics (TOG)*, 41(6):1–15, 2022.

[10] S. Bell, P. Upchurch, N. Snavely, and K. Bala. Material recognition in the wild with the materials in context database. *Computer Vision and Pattern Recognition (CVPR)*, pages 3479–3487, 2015.

[11] M. A. Lee, Y. Zhu, K. Srinivasan, P. Shah, S. Savarese, L. Fei-Fei, A. Garg, and J. Bohg. Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 8943–8950, 2019. doi:10.1109/ICRA.2019.8793485.

[12] Y. Gao, L. A. Hendricks, K. J. Kuchenbecker, and T. Darrell. Deep learning for tactile understanding from visual and haptic data. In *2016 IEEE international conference on robotics and automation (ICRA)*, pages 536–543. IEEE, 2016.

[13] X. Zhou, S. Lan, W. Wang, X. Li, S. Zhou, and H. Yang. Visual-haptic-kinesthetic object recognition with multimodal transformer. In *International Conference on Artificial Neural Networks*, pages 233–245. Springer, 2023.

[14] S. Luo, J. Bimbo, R. Dahiya, and H. Liu. Robotic tactile perception of object properties: A review. *Mechatronics*, 48:54–67, 2017.

[15] H. Bai, H. Chen, E. Healy, C. C. Kemp, and T. Bhattacharjee. Analyzing material recognition performance of thermal tactile sensing using a large materials database and a real robot. In *2022 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pages 2255–2262. IEEE, 2022.

[16] T. Bhattacharjee, J. Wade, Y. Chitalia, and C. C. Kemp. Data-driven thermal recognition of contact with people and objects. In *2016 IEEE Haptics Symposium (HAPTICS)*, pages 297–304. IEEE, 2016.

[17] M. Bednarek, M. R. Nowicki, and K. Walas. Haptr2: Improved haptic transformer for legged robots' terrain classification. *Robotics and Autonomous Systems*, 158:104236, 2022.

[18] C. Liu, H. Liu, H. Chen, W. Du, and H. Yang. Touchformer: A transformer-based two-tower architecture for tactile temporal signal classification. *IEEE Transactions on Haptics*, 2023.

[19] A. Kirillov, E. Mintun, N. Ravi, H. Mao, J. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

[20] S. Liu, Y. Lu, W. Zhang, X. Zhao, K. Wang, X. Bai, L. Zhang, B. Cai, J. Li, and X. Qi. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

[21] G. Zerveas, S. Jayaraman, R. Patel, A. Bhamidipaty, and C. Eickhoff. A transformer-based framework for multivariate time series representation learning. *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pages 647–656, 2021.

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pages 5998–6008, 2017.

[23] H. I. Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P.-A. Muller. Inceptiontime: Finding alexnet for time series classification. *Data Mining and Knowledge Discovery*, 34:1936–1962, 2020.