Cross-Modal Knowledge Distillation for Efficient Material Recognition: Aligning Language Descriptions with Tactile Image Models

Mashood M. Mohsan[1], Binzhao Xu[1], Basma B. Hasanen[1], Taimur Hassan[2], and Irfan Hussain*[1]

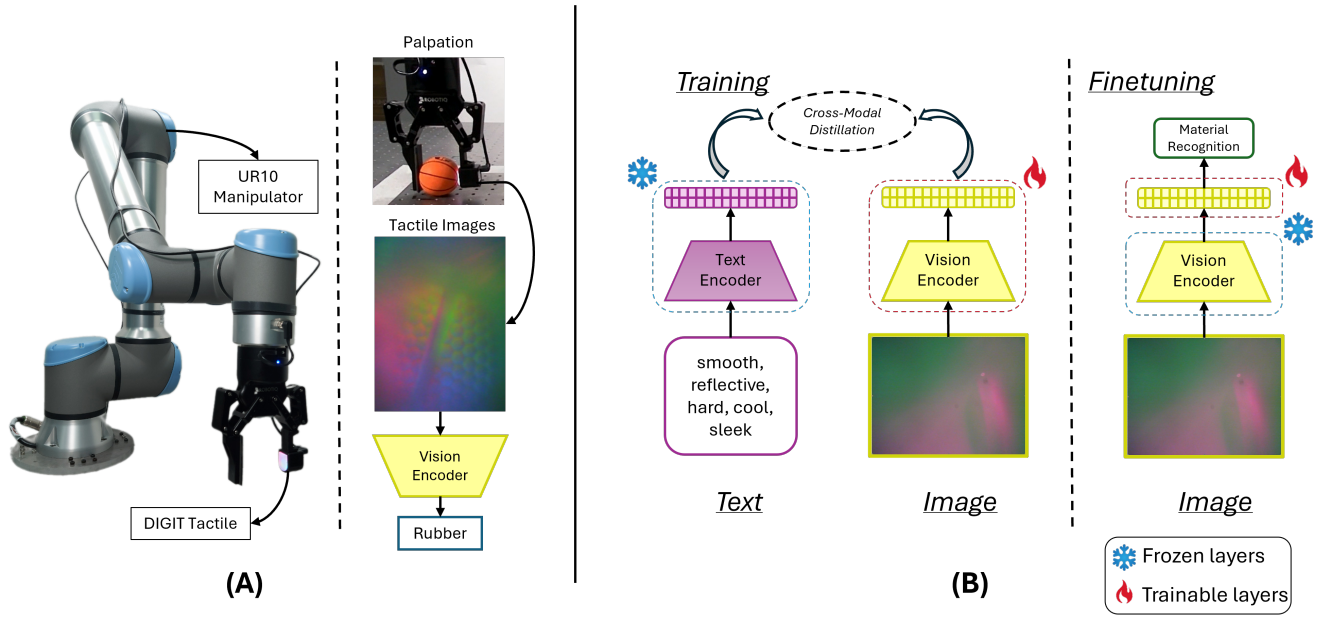

Fig. 1: **(A) Real-World Experiment Setup** The UR10 manipulator with a DIGIT tactile sensor performs real-world material recognition by palpating objects (e.g., a rubber ball shown). The distilled student vision model processes tactile images from the sensor to identify material properties. **(B) Knowledge Distillation Training and Finetuning Process**. *Left:* In the training phase, the text encoder (teacher model) processes textual descriptions of tactile properties, while the vision encoder (student model) processes the corresponding tactile images. The knowledge Distillation method aligns the student model's outputs with the teacher's, ensuring the student learns tactile representations based on the teacher model's fixed knowledge. *Right:* In the finetuning phase, the vision encoder is fine-tuned on a small number of samples to specialize in material recognition tasks based on the tactile images. The Encoder of Vision is frozen; only the classifier layer is fine-tuned.

***Abstract*— Material recognition is critical in robotics and automation, enabling systems to accurately identify and classify materials for tasks like manipulation and sorting. In this paper, we introduce a novel approach that leverages cross-modal knowledge distillation, where a language-based teacher model distills knowledge into a vision-based student model trained on tactile images. Using the pre-trained Bidirectional Auto-Regressive Transformer (BART) model as the teacher, which processes language descriptions of tactile properties, and a Vision Transformer (ViT) as the student model, we align tactile and language representations through a knowledge distillation framework. Our distilled ViT model achieved significantly higher accuracy (74.70%) in material recognition compared to a non-distilled ViT model (57.83%), demonstrating the value of integrating language-based knowledge for enhanced tactile material recognition. We also perform real word experimentation UR10 manipulator performing material recognition task.**

*Crossponding Author
[1]Khalifa University Center for Autonomous Robotic Systems (KU-CARS), Khalifa University, Abu Dhabi, UAE
[2]Department of Electrical, Computer and Biomedical Engineering, Abu Dhabi University, UAE

I. INTRODUCTION

Material recognition, the process of identifying and classifying materials based on their physical properties, plays a critical role in robotics, manufacturing, and quality control applications. It is essential for tasks like sorting, quality assurance, and automated assembly, where the precise identification of materials can significantly improve operational efficiency. In robotics, tactile sensing combined with vision-based methods allows for a deeper understanding of object properties, which is necessary for tasks such as grasping and manipulation in both structured and unstructured environments [1].

Several studies have explored material recognition using various modalities. Early research focused on vision-based methods, which rely heavily on visual features such as texture and color[2]. However, these methods face limitations in accurately identifying materials that are visually similar but differ in tactile properties, such as smoothness, hardness, or elasticity. Recent advancements have introduced vision-tactile sensors like GelSight, which capture high-resolution

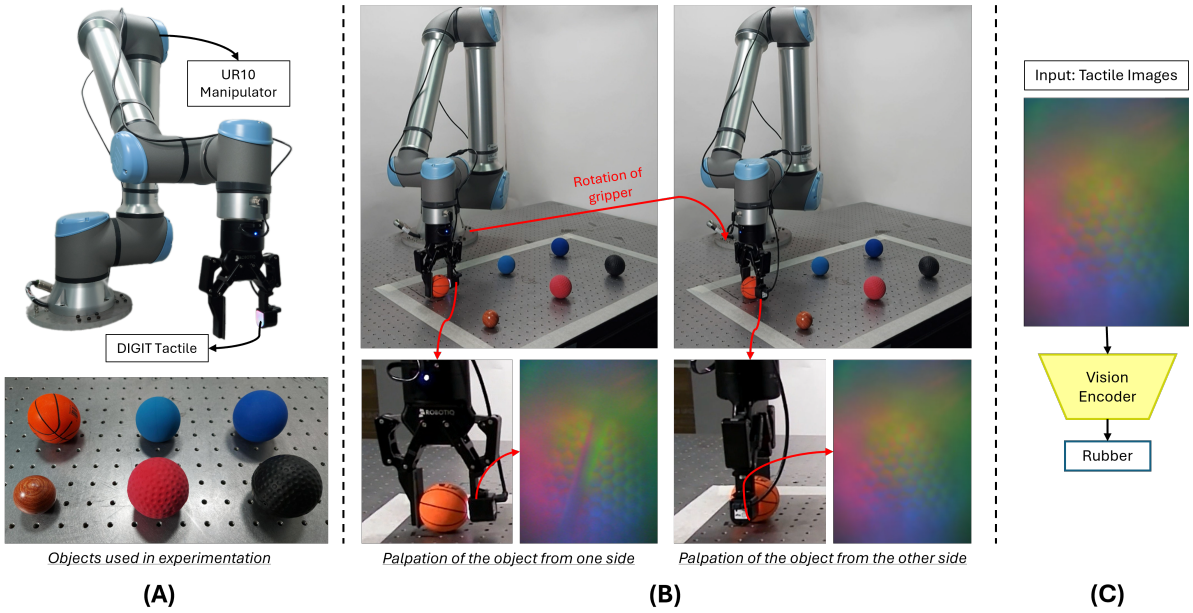

Fig. 2: Experimental Setup and Material Recognition Process **(A)** The experimental setup features a UR10 manipulator equipped with a DIGIT tactile sensor, used to palpate various objects made of rubber, plastic, wood, and silicone for material recognition. **(B)** Snapshots of the UR10 conducting tactile palpation on the objects. The UR10 performs palpation twice on each object, with a 90-degree gripper rotation before the second palpation to ensure material assessment. **(C)** The tactile images captured during palpation are fed into the distilled student model for material recognition, allowing the model to classify the object's material properties.

tactile information and have been successfully used for material classification. These sensors, combined with deep learning models, provide a more robust solution for material recognition, leveraging both visual and tactile data to improve classification accuracy[2].

Despite these advances, there remains a gap in fully leveraging multimodal learning for material recognition, particularly in the use of language-driven tactile classification. While vision and tactile sensors have been explored independently and jointly, few studies have investigated the potential of knowledge distillation methods, where language models guide the learning of tactile properties in vision-based systems. This research aims to bridge this gap by using a cross-domain knowledge distillation approach, transferring knowledge from language models to improve tactile material recognition performance.

The main contributions of this paper are:

- We propose a novel framework where a language-based teacher model distills knowledge into a vision-based student model for tactile material recognition.
- We created the first cross-modal dataset by modifying the TVL (Touch Vision Language) [3] dataset, which already included tactile images and language descriptions. We manually annotated material classes for approximately 49K pairs to support material recognition tasks.
- Real-world experimentation using UR10 to demonstrate the approach's effectiveness in practical material recognition tasks.

II. Methodology

Our approach consists of two main phases: knowledge distillation and fine-tuning. First, we applied cross-domain knowledge distillation to transfer knowledge from a pre-trained language model to a vision model. The language model processed textual descriptions of tactile properties, while the vision model learned to interpret corresponding tactile images. In the fine-tuning phase, the distilled vision model was further refined for material recognition tasks, focusing on classifying materials from only tactile images.

A. Cross Domain Knowledge Distillation

A knowledge distillation [4] method was employed in the training phase to transfer knowledge from a pre-trained language teacher model (BART) [5] to a vision-based student model (ViT)[6]. The aim was to enable the ViT model to interpret tactile properties from visual data. The teacher model processed textual descriptions of tactile properties, such as "soft," "fabric-like," "smooth," "reflective," and "hard," while its weights remained frozen to ensure consistent and stable guidance during training. The student model, ViT, received tactile images corresponding to these textual descriptions and was optimized through backpropagation. The KL Divergence loss function was used to align the student's output with the encoded knowledge of the teacher model, allowing the student to learn tactile representations aligned with the language-based model progressively.

Throughout the training process, only the student model's weights were updated, while the teacher's remained fixed. This approach allowed the student to internalize the tactile

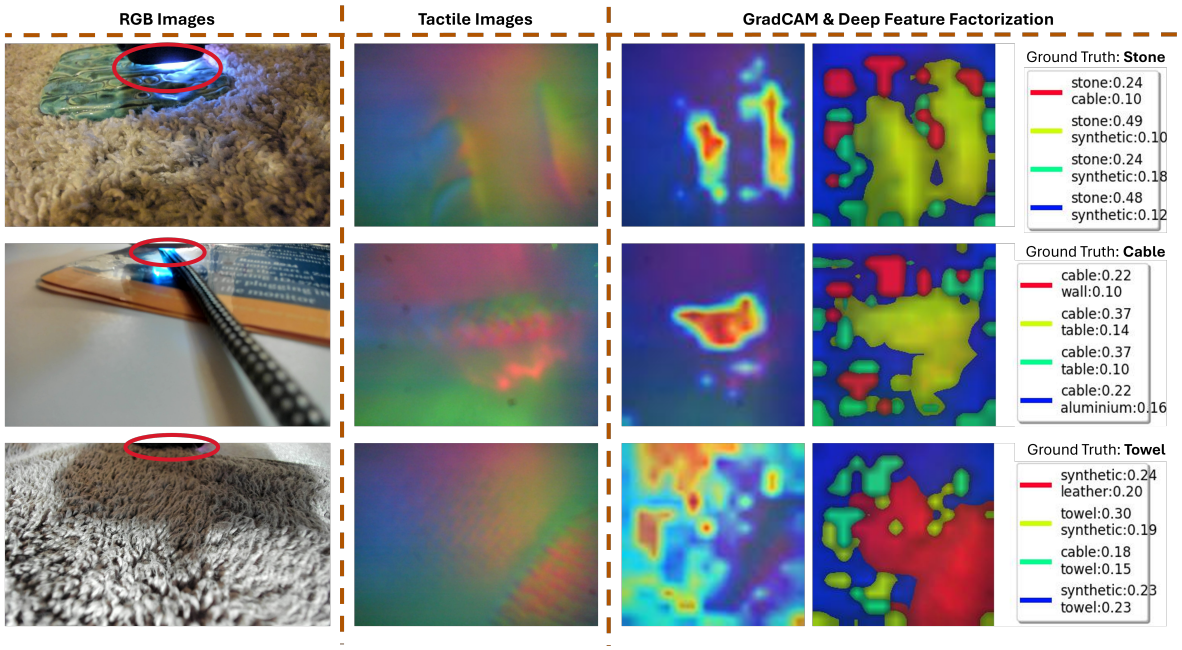

Fig. 3: The figure showcases RGB images (displayed for reference), tactile images (used by the model), and the corresponding GradCAM and deep feature factorization visualizations. The first two rows demonstrate successful material recognition where the classifier correctly identified the materials (Stone & Cable). In the third row, the classifier incorrectly predicted "Synthetic" instead of the correct material "Towel." The color-coded legend indicates the probability of each material class, highlighting the areas of interest that influenced the model's predictions.

properties conveyed by the language model. The iterative optimization ensured that the student could accurately generate tactile features from visual data, preparing it for the material recognition tasks in the subsequent phase.

B. Material Recognition

In the fine-tuning phase, the distilled ViT student model, pre-trained to interpret tactile features through knowledge distillation, was further refined for material recognition. This phase involved processing new tactile images and classifying materials. The vision encoder's layers, trained during the distillation phase, were kept frozen, and only the classifier head weights were updated to specialize the model for material recognition. The model was trained on additional labeled tactile image data, improving its performance in distinguishing between various materials in real-world scenarios. This fine-tuning allowed the model to leverage the distilled knowledge from the language domain while excelling in material classification tasks.

III. Experimentation Setup

A. Implementation Details

The proposed model was implemented using PyTorch and HuggingFace [7] on an NVIDIA RTX 4090 GPU with CUDA Toolkit v11.0.221 and cuDNN v7.5. To ensure training stability, some weights were initialized using pre-trained weights. Hyperparameters, including a learning rate of 0.00005, batch size of 64, and 50 epochs, were empirically determined. Accuracy was used as the primary metric for evaluating material recognition performance across all experiments.

B. Dataset Details

We trained our model using the TVL (Touch Vision Language) dataset [3], which includes 48,993 pairs of tactile images and text descriptions. Missing class information was manually annotated into material categories, and details are provided in Table I. The dataset was split 99%-1% for training and testing.

TABLE I: The table shows the dataset splits used for both the Knowledge Distillation (KD) phase and material recognition finetuning, detailing image-text pairs and training/testing splits for each material class.

Sr#	Type	Classes	Train	Test
1	*KD Phase*	Plastic	11176	103
2		Steel	6439	68
3		Fabric	4876	39
4		Metal	2668	20
5		Rubber	1814	20
6	*Material Recognition*	Synthetic	1615	14
7		Towel	1214	12
8		Carpet	790	10
9		Wood	715	9
10		Cable	713	8
11		Table	644	6
12		Wall	638	6
13		Polyester	527	5
14		Paper	492	4
15		Leather	411	3
16		Stone	354	3
17		Aluminum	249	2
18		Woven	238	1
Total			47975	1018

TABLE II: The table shows the accuracy of the distilled ViT student model on the fine-tuned dataset split, using different teacher models in the KD (Logits based) distillation process.

Sr#	Knowledge Distillation Teacher	Student	Accuracy
1	Roberta	ViT base	73.47
2	BART	ViT base	**74.70**
3	DistilBERT	ViT base	62.65

IV. Ablation Studies

A. Effect of various LLM Teachers

In the first ablation study, different teacher models were evaluated for their effectiveness in guiding the ViT student model. BART performed best, likely due to its architecture combining bidirectional encoding with a balanced model complexity that supports effective knowledge transfer during distillation. BART's design allows it to handle sequential dependencies and context effectively, crucial for transferring nuanced language representations to the vision-based ViT model. RoBERTa [8], while also capable of capturing rich language features, may have introduced too much complexity, making it harder for the student to generalize efficiently from the distilled information. DistilBERT [9], on the other hand, is optimized for reduced computational load, but this reduction in complexity may result in less detailed representations, which could explain the lower accuracy in guiding the ViT model. Results are shown in Table II.

B. Effect of various Learning Approaches

In this ablation study, we explored different distillation methods using the BART teacher and ViT student models, with Feature-Based Knowledge Distillation achieving the highest accuracy. This method's effectiveness stems from its direct transfer of intermediate feature representations, enabling the student model to align closely with the teacher's internal states. Logits-Based Knowledge Distillation also performed well by leveraging Kullback-Leibler divergence to align output distributions. In contrast, Decoupled Knowledge Distillation (DKD) [10] and Cosine Minimization underperformed, likely due to their inability to capture the full complexity of the feature space, either by separating key elements (DKD) or focusing too narrowly on vector alignment (Cosine Minimization). Results are shown in Table III.

TABLE III: Knowledge Distillation results using different distillation methods. The accuracy shown reflects the performance of the distilled ViT student model on the fine-tuned dataset split, using different distillation methods.

Sr#	Distillation Method	Model	Accuracy
1	KD (Feature Based)	BART+ViT	**89.16**
2	KD (Logits Based)	BART+ViT	74.70
3	Decouple KD (DKD)	BART+ViT	18.07
4	Cosine Minimization	BART+ViT	33.73
5	Non-distilled	ViT	57.83

V. Results

In the quantitative analysis, BART emerged as the best-performing teacher model in the knowledge distillation process, achieving a top accuracy of 74.70% when paired with the ViT base student model, as shown in Table II. Among the distillation methods, Feature-Based Knowledge Distillation produced the highest accuracy of 89.16%, as shown in Table III. Notably, the non-distilled ViT model achieved an accuracy of only 57.83%, underscoring the substantial improvement gained through language-based distillation. As depicted in Fig. 3, qualitative results further validate the model's performance. The GradCAM [11] and deep feature factorization visualizations highlight successful material recognition in the first two rows, where the classifier accurately identified materials like "Stone" and "Cable." However, a misclassification occurred in the third row, where "Synthetic" was predicted instead of "Towel," revealing areas for further improvement in tactile feature representation.

VI. Real World Experiment

The real-world experiment for material recognition was conducted using a UR10 manipulator equipped with a DIGIT [12] tactile sensor (Fig. 2). The UR10 was tasked with palpating various objects made from rubber, plastic, wood, and silicone. The objects (balls) were positioned in fixed locations throughout the experiment to ensure consistency. The UR10 applied a uniform force of 24N on each object during palpation to capture tactile images representing their material properties. Each object was palpated twice. Once from one side and again after rotating the gripper by 90 degrees to gather tactile information for more comprehensive material assessment (Fig. 2B). The goal was to use only tactile data to recognize and classify the objects based on their material properties.

VII. Conclusion

This work demonstrates the effectiveness of cross-modal knowledge distillation for material recognition using tactile images and language descriptions. Despite the promising results, there remains a significant need for larger and more diverse datasets to enhance generalizability. Future research should explore the integration of other vision-based models, such as CNNs or ResNet, to compare with or replace ViT to assess the versatility of our approach. Additionally, exploring more generalized models capable of performing across various material types and contexts is crucial for advancing tactile-based material recognition. Our results suggest that this distillation method has great potential but should be further validated across different datasets and architectures.

ACKNOWLEDGMENT

This publication is based upon work supported by the Khalifa University of Science and Technology under Award No RC1-2018-KUCARS.

REFERENCES

[1] R. Calandra, A. Owens, M. Upadhyaya, W. Yuan, J. Lin, E. H. Adelson, and S. Levine, "The feeling of success: Does touch sensing help predict grasp outcomes?" *arXiv preprint arXiv:1710.05512*, 2017.

[2] W. Yuan, Y. Mo, S. Wang, and E. H. Adelson, "Active clothing material perception using tactile sensing and deep learning," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 4842–4849.

[3] L. Fu, G. Datta, H. Huang, W. C.-H. Panitch, J. Drake, J. Ortiz, M. Mukadam, M. Lambeta, R. Calandra, and K. Goldberg, "A touch, vision, and language dataset for multimodal alignment," *arXiv preprint arXiv:2402.13232*, 2024.

[4] G. Hinton, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[5] M. Lewis, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *arXiv preprint arXiv:1910.13461*, 2019.

[6] D. Alexey, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv: 2010.11929*, 2020.

[7] T. Wolf, "Huggingface's transformers: State-of-the-art natural language processing," *arXiv preprint arXiv:1910.03771*, 2019.

[8] Y. Liu, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[9] V. Sanh, "Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[10] B. Zhao, Q. Cui, R. Song, Y. Qiu, and J. Liang, "Decoupled knowledge distillation," in *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 2022, pp. 11 953–11 962.

[11] J. Gildenblat and contributors, "Pytorch library for cam methods," https://github.com/jacobgil/pytorch-grad-cam, 2021.

[12] M. Lambeta, P.-W. Chou, S. Tian, B. Yang, B. Maloon, V. R. Most, D. Stroud, R. Santos, A. Byagowi, G. 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*, vol. 5, no. 3, pp. 3838–3845, 2020.