engineering proceedings

MDPI

Proceeding Paper

A Contrastive Learning Approach for Integrating Visuo-Tactile Representation in Textiles †

Nada Ech-chouqi [1,2] and Ghazal Rouhafzay [1,*]

[1] Department of Computer Science, Faculty of Sciences, University of Moncton, NB, E1A 3E9 Canada; echchouqinada@gmail.com
[2] École d'Ingénieurs du Littoral Côte d'Opale, CS 30613, 62228, Calais, France
* Correspondence: ghazal.rouhafzay@umoncton.ca
† Presented at The 11th International Electronic Conference on Sensors and Applications (ECSA-11), 26–28 November 2024; Available online: https://sciforum.net/event/ecsa-11.

Abstract: Vision and touch are fundamental sensory modalities that enable humans to perceive and interact with objects in their environment. Vision facilitates the perception of attributes such as shape, color, and texture from a distance, while touch provides detailed information at the contact level, including fine textures and material properties. Despite their distinct roles, the processing of visual and tactile information shares underlying similarities, presenting a unique opportunity to enhance artificial systems that integrate these modalities. However, existing methods for combining vision and touch often rely on data fusion at the decision level, requiring extensive labeled data and facing challenges in generalizing to novel situations. In this paper, we leverage contrastive learning to train a convolutional neural network on textile data using both visual and tactile inputs. Our objective is to develop a network capable of extracting unified representations from both modalities without the need for extensive labeled datasets. We explore using a contrastive loss functions to optimize the learning process. Our results demonstrate that the shared representations effectively capture critical data structures and features from both sensory modalities, enabling successful differentiation between object classes based on both vision and touch. We validate our approach through a series of experiments, optimizing hyperparameters to maximize performance. The findings suggest that extracting shared representations for vision and touch not only enhances the integration of visual and tactile information but also provides a robust framework for multimodal perception in artificial systems.

Keywords: robotics perception; deep learning; visuo tactile integration; contrastive learning

Citation: Ech-chouqi, N.; Rouhafzay, G. A Contrastive Learning Approach for Integrating Visuo-Tactile Representation in Textiles. *Eng. Proc.* **2024**, *5*, x. https://doi.org/10.3390/xxxxx

Academic Editor(s): Name

Published: 26 November 2024

1. Introduction

Vision and touch are two complementary sensing modalities that must synergistically collaborate to enable safe and effective object manipulation in robots. Efficient integration of these modalities is therefore a crucial step toward achieving full autonomy in robotic systems.

Inspired by how these senses complement each other in humans, researchers are working to enable robots with similar capabilities. However, training robots to understand touch, or haptic perception, presents challenges. Collecting large amounts of tactile data is time-consuming and resource-intensive, whereas visual data is far easier to gather at scale. This discrepancy motivates the need for innovative solutions that leverage visual data to enhance tactile understanding.

In this study, we focus on improving touch-based object recognition, specifically for fabric materials, by utilizing the strengths of visual data. Instead of relying solely on vast tactile datasets, we propose a contrastive learning framework that transfers knowledge from visual learning to touch-based recognition. By identifying shared representations

between these two modalities, such a method enables robots to understand tactile properties with minimal tactile data.

This approach can enhance robotic functions, such as identifying fabric textures, detecting surface irregularities, and recognizing material types based on visual input alone, significantly reducing the need for exhaustive tactile training. By finding common ground between vision and touch, we aim to create a more efficient and scalable solution for haptic perception in robotics. In the following sections, we present our framework and results, demonstrating how contrastive learning can effectively unify visual and tactile perception.

2. Literature Review

Object recognition through haptic perception is a growing area in robotics, despite challenges in collecting tactile data. Zhang et al. [1] improved object recognition using tactile data by optimizing a Support Vector Machine (SVM) with a Differential Evolution algorithm, enhancing accuracy and generalization. Other studies, like those by Jamali and Sammut [2] and Sugaiwa et al. [3], focused on material classification and grasp optimization. Jamali and Sammut classified materials by surface texture sensing, while Sugaiwa et al. developed a method for adjusting grasp force based on object properties.

Combining vision and touch for object recognition has been an interesting research topic. Early work by Allen [4] demonstrated the effectiveness of integrating visual and tactile feedback, where vision aids in object localization, and touch refines texture and shape understanding. Stansfield [5] expanded on this by developing a robotic system that uses passive vision for object detection and active touch for detailed analysis. More recent works such as the research by Yang and Lepora [6] introduce deep learning frameworks that allows robots to use both visual and tactile data for object recognition, particularly for irregular shapes. Rouhafzay and Cretu [7,8] further demonstrated the effectiveness of combining visual information as a guide to slectively collect tactile data for 3D object recognition.

Other works focus on enhancing the training of tactile object recognizers by leveraging visual information. Li et al. [9] employed generative models for cross-modal prediction, enabling robots to predict tactile inputs from visual data, enhancing perceptual accuracy. Rouhafzay et al. [10] proposed a hybrid deep learning model capable of handling both visual and tactile tasks by transfer learning from vision to touch. Yang et al. [11] and Lee et al. [12] further explored frameworks that integrate vision and touch, improving robot performance in contact-rich environments. Recently, Dave et al. [13] used a self-supervised contrastive approach to combine visual and tactile data, enhancing multimodal representation learning.

3. Framework

Figure 1 illustrates the overall framework of the proposed approach. We leverage a ResNet-50 architecture [14], pretrained on ImageNet [15], as the backbone and modify it to extract a shared representation from visual and tactile data. After feature extraction using ResNet-50 for both input modalities (visual and tactile), a custom projection head is added to transform these features into a more compact representation (128 dimensions) through dense layers. This approach is well-suited for contrastive learning scenarios, where the goal is to compare and learn representations from both tactile and visual inputs.

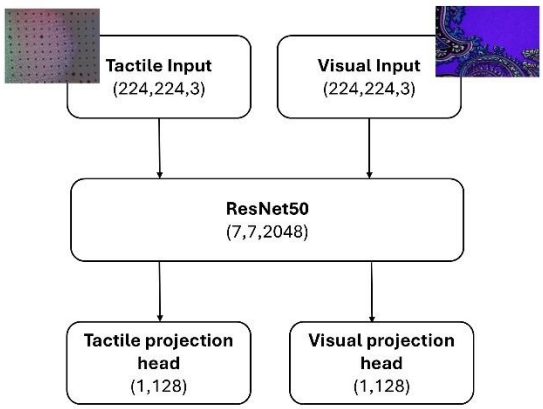


Figure 1. Overall framework to extract shared representation.

The contrastive loss function is used to bring the representations of similar samples from both modalities closer together, while pushing apart those of dissimilar samples. To compute the loss between the compact representations of the two modalities, we firstly normalize the projections for each modality as descrived by Equation (1).

$$z_i = \frac{p_i}{\|p_i\|_2}, \quad z_j = \frac{p_j}{\|p_j\|_2} \tag{1}$$

where p_i and p_j are the projections from the two modalities (visual and tactile), and $\|p_i\|_2$ and $\|p_j\|_2$ are the L_2 norm of these latent space vectors. Subsequently the cosine similarity is calculates as:

$$Similarity_{i,j} = \frac{z_i \cdot z_j}{\tau} \tag{2}$$

where τ is the temperature hyperparameter controling the model's sensitivity to differences between the pairs of projections. The lower the temperature, the more strongly the model penalizes the differences.

The labels are then created to represent positive pairs, that is, those where both inputs (projections1 and projections2) come from the same type of fabric, one from vision and the other from touch.

$$Labels_{i,j} = \begin{cases} 1, & if\ i = j \\ 0, & otherwise \end{cases} \tag{3}$$

To train the network we use cross-entropy loss to maximize the cosine similarity between the representations of data from vision and touch. As such, the network learns to extract a common representation between the two modalities.

4. Data and Experiemnt Setup

For our experiments, we used the ViTac [16] dataset and selected 11 fabric classes to simplify the training process. The visual data was captured using a Canon EOS Rebel T2i camera, which features an 18-megapixel APS-C CMOS sensor, providing high-resolution images (5184 × 3456 pixels) that highlight the texture details of the fabrics. The camera's 9-point autofocus system ensured sharp focus on the objects, while its DIGIC 4 processor facilitated fast image processing. The tactile data was collected using a GelSight [17] sensor, which provides detailed 3D images of fabric surfaces through contact with a soft elastomer membrane. This membrane, deformed by the object's texture, is illuminated by RGB LEDs, and the sensor captures high-resolution tactile information on shape, texture, and contact forces, enabling a rich perception of the fabrics. This setup allowed us to train a model with comprehensive visual and tactile data.

5. Results and Discussion

Temperature is a crucial hyperparameter in contrastive learning, particularly in regulating the sensitivity of the loss function to dissimilar objects. It directly influences how the model penalizes difficult negative samples—those fabric pairs that are challenging for the model to differentiate. A higher temperature reduces the sensitivity to these difficult negatives, resulting in less penalty for mistakes, while a lower temperature intensifies the penalty. Both extremes can negatively affect the model's performance, underscoring the importance of finding an optimal temperature. To determine the optimal temperature for our model, we conducted experiments by carefully adjusting the temperature value and analyzing the resulting performance.

Our results, visualized in Figure 2, indicate that a temperature of 0.08 produces the lowest average loss, making it the optimal value for our framework.

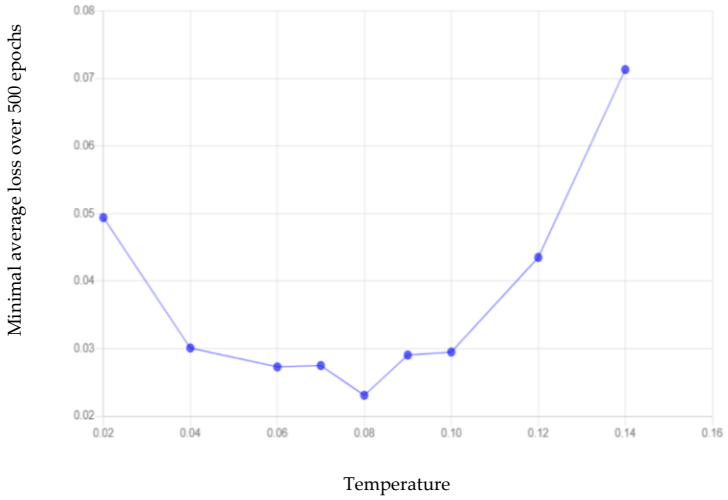


Figure 2. Comparison of the performance of models trained with different temperature values.

After determining the optimal temperature, we visualized the data using t-SNE [18] to better understand how well the model could differentiate between fabric types across both visual and tactile data. This plot allows us to represent data, which are modeled in multiple dimensions, in a two- or three-dimensional graph while preserving the structure of the data. The axes of this graph do not represent specific characteristics of the data but rather correspond to artificial dimensions we refer to as 'components'. Here we leverage a two-dimensional representation for both tactile and visual data. In the Figure 3, the numbers in the legend represent the tissue labels (with 11 representing tissue '11'), so each color corresponds to a specific tissue. The goal of this project is to recognize objects, which is reflected in the formation of clusters of identical colors, separated from one another.

As previously mentioned, in contrastive learning, the temperature parameter plays a crucial role in adjusting the similarity scores used to compute the loss, effectively controlling the sharpness of the similarity distribution. A higher temperature value encourages the model to consider a broader range of training examples for feature extraction, while a lower temperature sharpens the distribution, increasing the contrast between similar and dissimilar pairs.

Upon analyzing the results for different temperature values, it is clear that, especially in the case of visual data, a lower temperature leads to more distinct visual representations between different classes, as shown by the t-SNE. This result is expected, as visual data inherently encompasses greater variability, making lower temperatures more effective for extracting discriminative features.

Conversely, tactile data is generally associated with higher levels of noise, suggesting that a higher temperature could be beneficial in reducing the network's sensitivity to

noise. Thus, when working with visuo-tactile image pairs, finding an optimal temperature value that balances both modalities is essential.

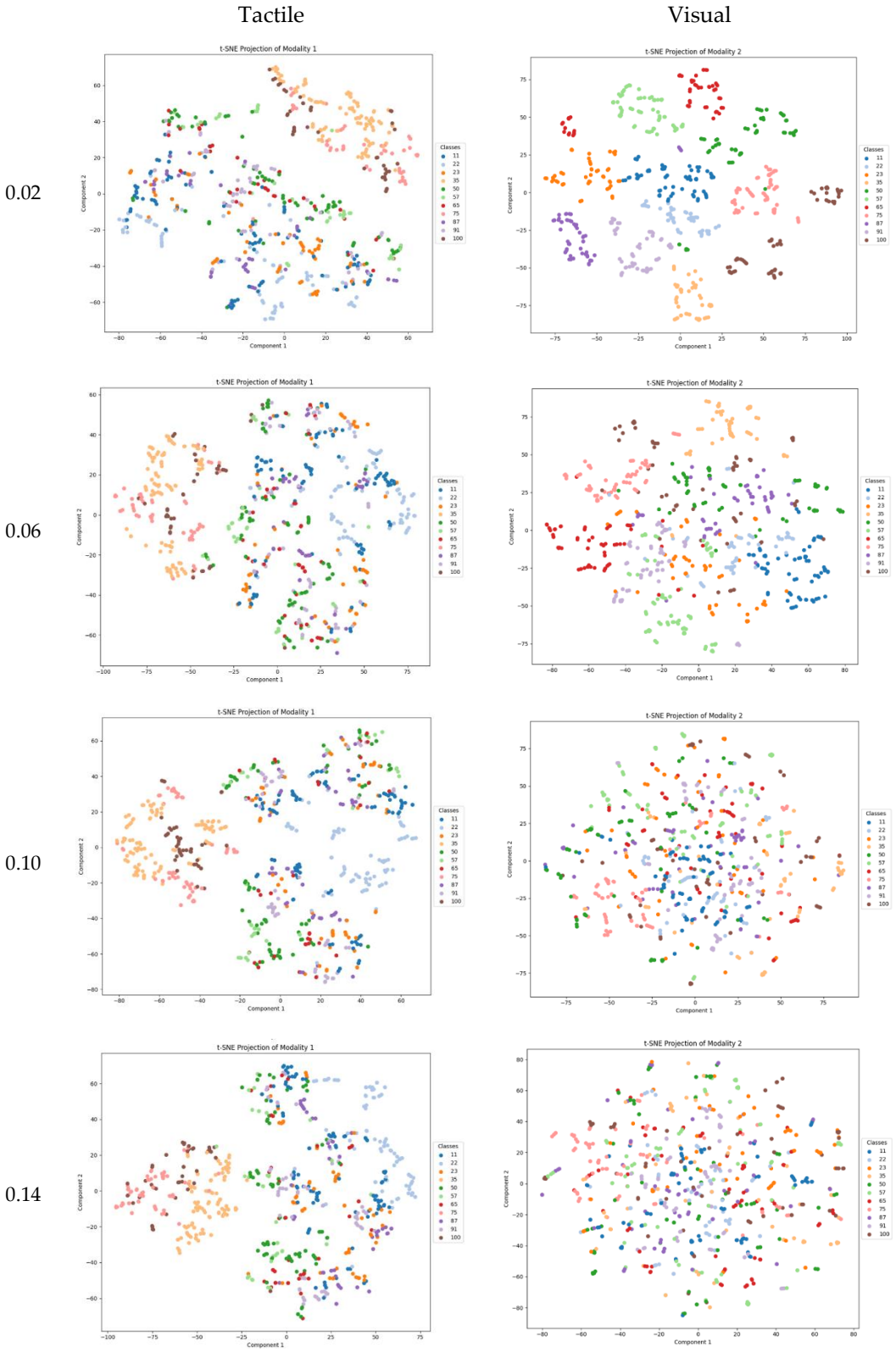
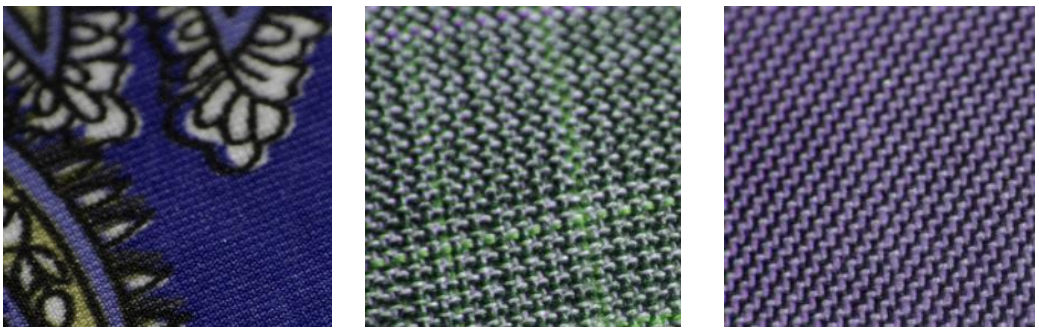


Figure 3. Comparison of the performance of models trained with different temperatures.

Furthermore, the t-SNE visualizations of the tactile data indicate that classes 23, 65, and 91 exhibit the greatest dispersion of features, making it difficult for the network to effectively discriminate between them. A visual inspection of these classes, as illustrated in Figure 4, confirms that their textures are quite similar. This limitation is primarily due to the constraints of the tactile sensor rather than the quality of the learned tactile representations.



(**a**) (**b**) (**c**)

Figure 4. (**a**) visual image of tissue number 23; (**b**) visual image of tissue number 65; (**c**) visual image of tissue number 91.

6. Conclusion

In this paper, we presented a contrastive learning framework designed to extract shared representations from visual and tactile data of various fabrics. Unlike conventional approaches that pair an image with its augmented versions in transfer learning, we paired visual images of an object with their corresponding tactile images. This strategy allowed us to establish meaningful connections between the two modalities.

To assess the discriminative power of the extracted features across different fabric types for both vision and touch, we used t-SNE visualizations. These results confirmed that the shared representation successfully enables object classification through both sensory modalities, demonstrating the effectiveness of our approach in bridging visual and tactile perception.

Author Contributions: Conceptualization, G.R.; methodology, G.R. and N.E.; software, G.R. and N.E.; validation, G.R. and N.E.; formal analysis, G.R. and N.E.; investigation, G.R. and N.E.; resources, G.R. and N.E.; data curation, G.R. and N.E.; writing—original draft preparation, N.E.; writing—review and editing, G.R.; visualization, G.R. and N.E.; supervision, G.R.; project administration, G.R.; funding acquisition, G.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Research NB Talent Recruitment Fund.

Institutional Review Board Statement: Not applicable

Informed Consent Statement: Not applicable

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zhang, X.; Li, S.; Yang, J.; Wang, Y.; Huang, Z.; Zhang, J. Tactile Perception Object Recognition Based on an Improved Support Vector Machine; *Micromachines* **2022**, *13*, 1538. https://doi.org/10.3390/mi13091538.
2. Jamali, N.; Sammut, C. Material Classification by Tactile Sensing Using Surface Textures. In Proceedings of the 2010 IEEE International Conference on Robotics and Automation (ICRA); Anchorage, AK, USA, 3–8 May 2010; pp. 2336–2341. https://doi.org/10.1109/ROBOT.2010.5509675.

3. Sugaiwa, T.; Fujii, G.; Iwata, H.; Sugano, S. A Methodology for Setting Grasping Force for Picking up an Object with Unknown Weight, Friction, and Stiffness. In Proceedings of the 2010 10th IEEE-RAS International Conference on Humanoid Robots (Humanoids), Nashville, TN, USA, 6–8 December 2010; pp. 288–293. https://doi.org/10.1109/ICHR.2010.5686331.
4. Allen, P.K. Integrating Vision and Touch for Object Recognition Tasks. *Int. J. Robot. Res.* **1988**, *7*, 15–33. https://doi.org/10.1177/02783649880070060.
5. Stansfield, S.A. A Robotic Perceptual System Utilizing Passive Vision and Active Touch. *Int. J. Robot. Res.* **1988**, *7*, 138–161. https://doi.org/10.1177/02783649880070061.
6. Yang, C.; Lepora, N.F. Object Exploration Using Vision and Active Touch. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Vancouver, CO, Canada, 24–28 September 2017; pp. 138–145. https://doi.org/10.1109/IROS.2017.8206542.
7. Rouhafzay, G.; Cretu, A.M. Object Recognition from Haptic Glance at Visually Salient Locations. *IEEE Trans. Instrum. Meas.* **2019**, *69*, 672–682. https://doi.org/10.1109/TIM.2019.2930389.
8. Rouhafzay, G.; Cretu, A.M. An Application of Deep Learning to Tactile Data for Object Recognition under Visual Guidance. *Sensors* **2019**, *19*, 1534. https://doi.org/10.3390/s19071534.
9. Li, Y.; Zhu, J.; Tedrake, R.; Torralba, A. Connecting Touch and Vision via Cross-Modal Prediction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019. https://doi.org/10.48550/arXiv.1906.06322.
10. Rouhafzay, G.; Cretu, A.M.; Payeur, P. Transfer of Learning from Vision to Touch: A Hybrid Deep Convolutional Neural Network for Visuo-Tactile 3D Object Recognition. *Sensors* **2021**, *21*, 113. https://doi.org/10.3390/s21010113.
11. Yang, F.; Feng, C.; Chen, Z.; Park, H.; Wang, D.; Dou, Y.; Zeng, Z.; Chen, X.; Gangopadhyay, R.; Owens, A.; Wong, A. Binding Touch to Everything: Learning Unified Multimodal Tactile Representations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–20 June 2024. https://doi.org/10.48550/arXiv.2401.18084.
12. Lee, M.A.; Zhu, Y.; Zachares, P.; Tan, M.; Srinivasan, K.; Savarese, S.; Li, F.; Garg, A.; Bohg, J. Making Sense of Vision and Touch: Learning Multimodal Representations for Contact-Rich Tasks. *IEEE Trans. Robot.* **2020**, *36*, 582–596. https://doi.org/10.48550/arXiv.1907.13098.
13. Dave, V.; Lygerakis, F.; Rueckert, E. Multimodal Visual-Tactile Representation Learning through Self-Supervised Contrastive Pre-Training. *arXiv* **2024**, arXiv:2401.12024. https://doi.org/10.48550/arXiv.2401.12024.
14. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
15. Deng, J.; Dong, W.; Socher, R.; Li, L.; Li, K.; Fei-Fei, L. ImageNet: A Large-Scale Hierarchical Image Database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Miami, FL, USA, 20–25 June 2009; pp. 248–255.
16. Luo, S.; Yuan, W.; Adelson, E.; Cohn, A.G.; Fuentes, R. ViTac: Feature Sharing between Vision and Tactile Sensing for Cloth Texture Recognition. In Proceedings of the 2018 IEEE International Conference on Robotics and Automation (ICRA), Brisbane, Australia, 21–25 May 2018; pp. 2722–2727.
17. Yuan, W.; Dong, S.; Adelson, E.H. GelSight: High-Resolution Robot Tactile Sensors for Estimating Geometry and Force. *Sensors* **2017**, *17*, 2762. https://doi.org/10.3390/s17122762.
18. Van der Maaten, L.; Hinton, G. Visualizing Data Using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.