

Investigating Vision Foundational Models for Tactile Representation Learning

Ben Zandonati*

University of Cambridge
baz23@cam.ac.uk

Ruohan Wang*

Institute for Infocomm Research, A*STAR
john.rh.wang@gmail.com

Ruihan Gao

Carnegie Mellon University
ruihang@andrew.cmu.edu

Yan Wu

Institute for Infocomm Research, A*STAR
wuy@i2r.a-star.edu.sg

Abstract—Tactile representation learning (TRL) equips robots with the ability to leverage touch information, boosting performance in tasks such as environment perception and object manipulation. However, the heterogeneity of tactile sensors results in many sensor- and task-specific learning approaches. This limits the efficacy of existing tactile datasets, and the subsequent generalisability of any learning outcome. In this work, we investigate the applicability of vision foundational models to sensor-agnostic TRL, via a simple yet effective transformation technique to feed the heterogeneous sensor readouts into the model. Our approach recasts TRL as a computer vision (CV) problem, which permits the application of various CV techniques for tackling TRL-specific challenges. We evaluate our approach on multiple benchmark tasks, using datasets collected from four different tactile sensors. Empirically, we demonstrate significant improvements in task performance, model robustness, as well as cross-sensor and cross-task knowledge transferability with limited data requirements.

I. INTRODUCTION

The sense of touch allows humans to feel, understand and ultimately manipulate through physical interaction. It is vital for exploration, object discrimination and fine-grained control, especially where visual perception lacks the resolution to detect surface changes, or is denied entirely. Inspired by the human sense of touch, robotic tactile learning has improved performance in tasks such as object/environment recognition [1, 2], pick-and-place [3] and in-hand manipulation [4].

Tactile representation learning (TRL) leverages machine learning (ML) to make sense of the rich data generated by specialized tactile sensors. Design choices such as sampling resolution, operating conditions and cost result in different tactile sensors adopting distinct sensing mechanisms (e.g. visual signals [5] and barometric signals [6]). Ideally, TRL should be sensor-agnostic, accommodating various data formats of different sensors and able to construct consistent representations of objects and environments. In practice, however, most methods developed are sensor-specific with tailored architectures and data processing routines [e.g. 5, 7, 8, 9].

This siloed approach has multiple limitations. First, individual tactile datasets are usually small due to the high cost of data collection. The tactile representation derived from such small datasets often generalize less well, especially for out-of-distribution data [e.g., 8, 10]. Even calibration differences and expected wear from regular usage present domain shifts detrimental to model performance. Furthermore, the lack of a unifying data format for different tactile sensors makes it

difficult to reuse knowledge captured in learned representations. For a new sensor design, the accompanying tactile representation model has to be learned from scratch, along with expensive data collection. All these limit the effectiveness and efficiency of TRL.

The above limitations are further highlighted when we contrast TRL with other application domains like computer vision (CV), and natural language processing (NLP). Both CV and NLP benefit from a unifying input format (images and text respectively), which permits fully shared model architectures for convenient knowledge transfer. In particular, foundational models [11] are trained on massive datasets such as ImageNet [12] and CommonCrawl [13] to derive general representational knowledge, which can be specialized to diverse downstream tasks, such as semantic segmentation [14] in CV, and sentiment analysis [15] in NLP. Foundational models improves learning efficiency and model robustness of downstream tasks, especially for limited training data [15].

Biologically, the human somatosensory system shares similar neural mechanisms with the visual cortex responsible for processing spatial features [16]. This implies that tactile properties such as texture are largely descriptions of surface spatial properties [17], motivating the question of whether *a vision foundational model could be exploited to tackle the aforementioned challenges in TRL*. Specifically, we investigate the following:

- Can vision models be agnostic to data from heterogeneous tactile sensors?
- Can vision foundational models improve model performance and robustness for TRL?
- Can vision architecture facilitate efficient knowledge transfer between downstream learning tasks and models trained on different sensor data?

In this work, we present a unified approach to address the above questions. We first present the use of *tactile images* as a simple unifying data format for heterogeneous tactile sensory outputs, to encode them as spatial features. This recasts TRL as a vision task, but with different input image sizes for different sensors. We adopt convolutional models [18] as the fully shared architecture for all sensors, exploiting convolution’s agnosticity to image sizes.

The above construct enables efficient knowledge transfer in multiple ways. First, we show that a foundational vision model pre-trained on natural images can be directly applied to tactile

learning tasks by simply performing least square regression to the last layer, providing evidence on the connection between visual and tactile perception in a non-biological system. Second, the foundational model can also be fine-tuned into tactile representation models with improved performance and robustness. In particular, we leverage data augmentation to counteract the limited tactile data during fine-tuning. Lastly, we demonstrate that the fine-tuned tactile representation model retains general features to allow cross-task and cross-sensor transfer.

To evaluate our proposed approach, we consider multiple benchmark tasks including standard material classification, continual learning for material classification and detection of fabric composition. We specifically test on data collected from four different sensors, with different data collection procedures, to demonstrate the general applicability of our approach.

Contributions. Our key contributions are summarized below:

- We extensively investigate on the feasibility, effectiveness, efficiency and robustness of using a vision foundational model for TRL. We use tactile images as a unified model input transformed from any tactile sensors.
- We introduce a new evaluation benchmarks for tactile learning, namely fabric composition detection.
- We contribute two new tactile datasets, including a material classification dataset using GelSight sensor and a fabric composition dataset using Contactile sensor.
- Empirically, we demonstrate that our proposed approach learns robust models for all sensors evaluated and outperforms baseline models tailored to specific sensors.

II. PRELIMINARIES AND RELATED WORK

We present three task settings to support the comprehensive evaluation of our proposed approach. The first two tasks are standard benchmarks for TRL while the third one is a novel task of composition detection task. We also review relevant works.

A. Tactile Representation Learning Tasks

Material Classification. This is a common benchmark for TRL [e.g. 8, 19, 20, 21, 22, 23, 24]. Similar to image classification, material classification determines the source material measured by a tactile sensor, from a finite number of classes. For example, early research involved classification of the textural information gathered via sliding an electret microphone across the surface of materials [25]. The task remains a standard benchmark amid the rapid development of different sensor designs.

A natural extension to standard material classification investigates the learned model’s robustness to out-of-distribution data. This includes varying data length and the moving speed of the tactile sensor (as controlled by a robot). For example, [26] achieved improved robustness to the sensor’s movement speed via additional sensing modalities. [8] also proposed a customized spiking neural network to reduce the data length needed for classification.

Continual Learning for New Materials. For real-world applications, robots are expected to continuously learn and adapt to novel environments. This also applies to TRL and was investigated in [27, 28], where robots learn new objects continuously by touch. In this work, we similarly extend material classification to the continual learning (CL) [29] setting. Formally, let $D = \{B_1, B_2, \dots, B_T\}$ be a data sequence with B_t denoting the data for material t . We wish to design a CL algorithm $\text{Alg}(\cdot)$ in the form of

$$(f_t, M_t) = \text{Alg}(B_t, f_{t-1}, M_{t-1}), \quad (1)$$

where f_t is the current classification model after learning the novel material t . f_t should be capable of classifying all materials observed so far (i.e., B_1 through B_t). A small memory buffer M is allowed to store data about previous materials to mitigate model forgetting. M_t denotes the current content of the memory buffer.

Intuitively, the CL algorithm $\text{Alg}(\cdot)$ must learn each material sequentially. It also cannot access training data for previous materials except for those stored in the memory buffer. The algorithm is thus forced to learn new materials on the fly without forgetting its existing knowledge. In contrast, standard material classification learns all materials in D concurrently and with unlimited access to all data. CL thus represents a more challenging and realistic benchmark.

Fabric Composition Detection. We introduce a new evaluation benchmark for TRL. Concretely, we design a fine-grained fabric composition detection task, in which the learned tactile model must predict the constituents of a specific fabric material, instead of simply identifying it. This task serves as a more challenging benchmark compared to standard material classification. It also allows us to investigate knowledge transfer between sensors and tasks (e.g., from material classification to constituents detection). We will describe the new dataset collected for this task in Sec. III.

B. Existing Methods

There exists a wide range of tactile sensor designs leveraging various sensing modalities, including strain gauges [24], piezo-resistive layers [30], accelerometers [31], capacitive [32], optical [5, 33] and those combining multiple sensing mechanisms [34, 35]. Most tactile learning methods tailor their respective model architectures and learning algorithms to the specific sensors used [e.g., 8, 23, 24, 36]. These existing approaches learn sensor-specific mappings from raw sensor output to some latent representation, and adjust the model size based on size of sensor output. These tailored decisions inevitably lead to a siloed state for TRL: the developed models can’t be easily reused for different sensors, even when the desired ML task remains identical.

[10] partially addresses the above issues by learning a shared latent representation for two different sensors. This approach demonstrates improved performance compared to independently learning each sensor’s data. However, it must still learn sensor-specific mappings from raw data to the shared representation, thus limiting its reuse potential for additional

sensors. In contrast, our proposed approach standardises the transformation to map any raw sensor data to tactile images, to be processed by a fully shared ML model. As we will demonstrate in our experiments, our approach grants more flexibility towards knowledge transfer.

III. SENSORS AND DATASETS

We present the sensors and the associated datasets considered in this work. They are intended to validate the general applicability of our approach, and to contextualize the challenge posed by heterogeneous sensors. Each dataset is used for one or more learning tasks described in Sec. II-A.

RoboSkin. Roboskin is a capacitive sensor designed for iCub [32]. Taunyazov et al. [36] collected a material classification dataset using the RoboSkin on the iCub robot forearm, sweeping across multiple materials without strict control of velocity and exerted forces. This public dataset contains 20 different materials with 50 samples in each class. Each sample contains 75 sensor readings.

BioTac. SynTouch BioTac® is a multi-modal tactile sensor using fluid pressure sensor and thermistor [37]. Gao et al. [10] released a material classification dataset using the BioTac sensor fitted as an extended end-effector on a KUKA LBR iiwa 14 robot arm, sliding laterally across different materials with controlled speed and contact force. BioTac-20 dataset contains the same 20 materials as the RoboSkin dataset with 50 samples in each class. Each sample contains 400 readings. A larger BioTac-50 dataset was later released.

We contribute two new datasets using alternative sensors. We will release both datasets publicly to support future research in the community.

GelSight. Gelsight is a camera-based sensor producing images of the contact surface, showing surface geometry and deformation with a soft elastomer [5]. Each reading is an image of 480×640 . A material classification dataset consists of 45 materials with 50 samples in each class. As the elastomer is vulnerable to abrasion from sliding motion, data is collected by rolling the sensor locally on material surfaces. The sensor, mounted on a KUKA LBR iiwa 14 robot arm, touches the material surface from above with a 1N force threshold. The sensor is then rotated clockwise by 1 degree, anticlockwise by 2 degrees, and finally clockwise by 1 degree back to the centre position (illustrated in Fig. 1a).

Contactile. Contactile® sensor uses a soft, silicone array based on PapillArray [7]. The sensor measures deflection, force and vibration. We collect the data using two protocols. Protocol 1 is identical to that of BioTac dataset. In Protocol 2, the sensor is handheld and slid across materials casually with different contact forces, speeds and along different directions, to mimic more realistic and natural movements. The dataset contains samples collected from 32 fabrics, each consisting of possible 6 constituent materials: Linen, Viscose, Cotton, Wool, Polyester and Elastane (see Tab. I for examples). 40 and 10 samples per material are collected for Protocols 1 and 2 respectively. The collection setup is illustrated in Figs. 1b and 1c.

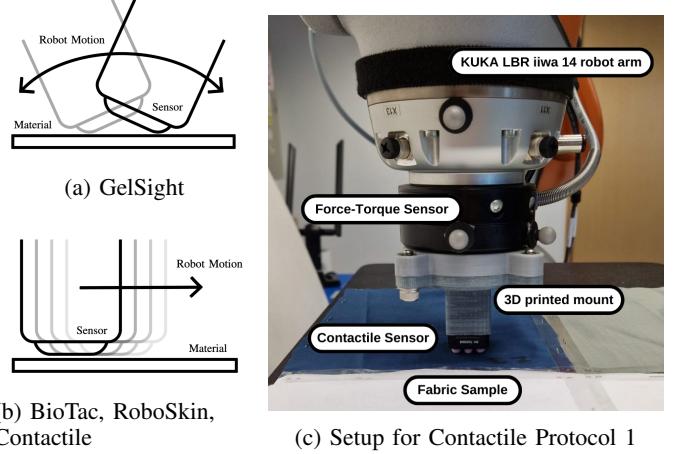


Fig. 1: (a) and (b) are illustrations of tactile data collection process for different sensors. (c) is the robot setup for Protocol 1 in Contactile data collection.

TABLE I: Fabric examples and their composition materials

Material	Image	% by mass				
		Linen	Viscose	Cotton	Wool	Polyester
Cotton-Linen		45	0	55	0	0
Poplin		0	0	20	0	80
Drill Stretch		0	0	100	0	0
Felt		0	65	0	35	0

IV. METHOD

We present a unified approach to tackle heterogeneous sensors and efficient knowledge transfer in TRL. Our approach relies on a unifying format for different sensor data, and exploits convolution’s agnosticity to input size to enable fully shared models. These fully shared models in turn enables convenient knowledge transfer. We also discuss data augmentations to counteract limited tactile training data. Lastly, we discuss a continual tactile learning approach as a direct application of knowledge transfer.

A. Tactile Images and Convolutional Architectures

We use simple transformations to convert data generated by various sensors into 2D images, which serves as the unified input format for the subsequent ML models. Specifically, tactile images aims to transform tactile sensory output into an encoding of the global geometry for the contact surface. This transformation is inspired by the processing similarities between the human visual cortex and somatosensory system [16], and captures the intuition that significant tactile properties are fundamentally spatial [17].

Camera-based sensors such as GelSight directly capture *global* surface geometry as images and can be used as model

input directly. However, non-camera-based sensors typically have sparse sensing points that only produce *localized* signals about the contact surface. To better encode the global surface geometry, we thus require more local samples that span across the contact surface. This could be conveniently achieved by concatenating consecutive vectors from the tactile data stream, as the sensor slides over the contact surface. Formally, let $S = \{s_1, s_2, \dots, s_T\}$ be the data stream produced by a sensor sliding across a surface, where $s_t \in \mathbb{R}^n$ is a single reading from the sensor. We define a tactile image as a matrix $\text{Im}(S) = [s_j, s_{j+1}, \dots, s_k]$ for some constant j, k . Intuitively, $\text{Im}(S)$ leverage the temporal dimension of tactile data stream to better encode global surface properties (see also Fig. 2 for an illustration).

We note that tactile images of different sensors still have different dimensions. To achieve fully shared models for knowledge transfer, we thus adopt convolutional architectures such as ResNet [38], since convolution does not require a fixed input size. ResNet is also a representative state-of-the-art model for processing spatial input, including the surface geometry encoded in tactile images.

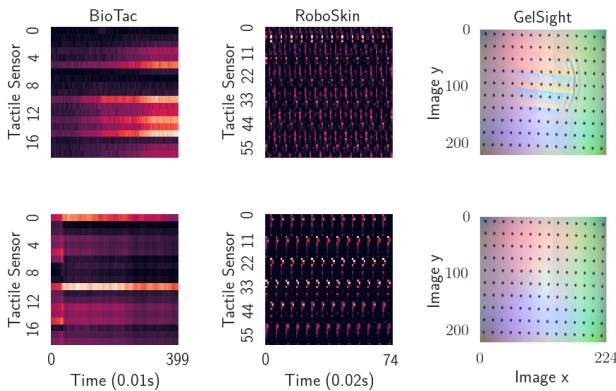


Fig. 3: Tactile image representations for the BioTac, RoboSkin and GelSight sensors for two material classes.

B. Model Training

With tactile images and our chosen model architecture, we effectively recast TRL as a vision task. For training, we minimise the empirical cross-entropy loss

$$\arg \min_f \sum_{(x,y) \in D} \ell_{ce}(f(x), y) \quad (2)$$

where f is the model and ℓ_{ce} is the cross-entropy loss. D denotes the dataset containing labeled tactile images (x, y) .

Crucially, we can initialize f with a pre-trained model to enable knowledge transfer. In particular, we may interpret TRL

as a downstream task for a vision foundational model on general spatial features. In our experiments, we will demonstrate that a foundational model trained on natural images already robustly encodes the general features required for tactile images.

Data Augmentation. As discussed earlier, tactile datasets are typically small due to the high cost of data collection due to the interactivity of the modality and significant wear and tear. Data augmentation is therefore important to mitigate model overfitting, especially for larger architectures like ResNet. We propose to directly apply standard CV augmentations: *resizing*, *cropping*, *flipping* and *jittering*. We observe that each of these augmentations encodes a meaningful variation to the data collection process, even for non-camera-based sensors. For instance, cropping the tactile images encodes varying the duration of robot motion during data collection. Tab. II lists all chosen augmentations and their interpretation.

TABLE II: Tactile images augmentations and their physical interpretation

Augmentation Technique	Physical Interpretation
Flipping (along data axis)	Reversing the direction of robot motion.
Resizing (along temporal axis)	Vary the speed of robot motion.
Cropping (along temporal axis)	Vary the duration of robot motion.
Jittering	Simulate sensor noise and drift.

The chosen augmentations are readily accessible from common deep learning frameworks [39] and may be directly applied. We will demonstrate empirically that the augmentations is crucial to model robustness.

C. Continual Tactile Learning

As robots are increasingly expected to work in unstructured environments, continual learning of unordered new percepts is important. Sec. II-A introduced continual learning (CL) of new materials as a natural extension to standard material classification. The two key challenges for CL are: 1) whether robots could learn about new materials on the fly, and 2) continuous learning does not cause catastrophic forgetting of current knowledge [40, 41].

We adopt schedule-robust online continual learning (SCROLL) [42] to tackle CL of new materials. We choose SCROLL because the method leverages pre-trained models for efficient knowledge transfer, thus allowing new materials to be learned with limited interaction. In addition, SCROLL is robust to the schedule under which the data is presented (e.g., the order in which each material is learned), a crucial property to ensure model reliability in real-world situations.

Using the notations introduced in Eq. (1), we characterize SCROLL as a two-phase process. Given a suitable pre-trained embedding model ψ , we first learn an online linear classifier ϕ_t via recursive least squares [43] as novel material data B_t is observed. We then fine-tune the composite model $f_t = \psi \circ \phi_t$ using the current memory buffer M_t to yield f_t^* . Both f_t and f_t^* are valid CL models for all data observed so far, with f_t^* having a fine-tuned representation based on the observed data.

SCROLL uses exemplar selection [44] for updating M_t . The overall algorithm is presented in Alg. 1,

Algorithm 1 SCROLL (incremental)

Initialization: Buffer $M_0 = \emptyset$, data statistics $c_y^0 = 0, A_0 = 0$
Input: Embedding model ψ , next data batch B_t , current buffer M_{t-1} , current data statistics c_y^{t-1}, A_{t-1}

$$c_y^t, A_t = \text{RecursiveLeastSquare}(c_y^{t-1}, A_{t-1})$$

$$\phi_t = \text{RidgeRegressor}(c_y^t, A_t)$$

$$f_t = \phi_t \circ \psi$$

$$M_t = \text{SelectExemplar}(M_{t-1}, B_t, \psi)$$

$$f_t^* = \text{FineTune}(f_t, M_t)$$
Return c_y^t, A_t, M_t, f_t and f_t^*

where c_y, A are necessary data statistics for recursive least squares (see [42] for further details on SCROLL).

V. EXPERIMENTS

We evaluate our approach extensively across a wide variety of sensors and tasks, as introduced in Sec. II and III. Our experiments address the following questions:

- Is our approach generally applicable to heterogeneous tasks and sensors? How does our approach compared to sensor-specific methods?
- What are the effects of tactile image augmentation?
- Does our approach allow efficient knowledge transfer? What are the effects of knowledge transfer?

Data Pre-Processing. Following Sec. IV-A, we transform BioTac data into 19×400 images by stacking 400 consecutive vectors. This corresponds to 4 seconds of data. RoboSkin data is transformed into 60×75 images, corresponding to 1.5 seconds of data. Lastly, Contactile data is transformed into 27×599 images, which is 6 seconds of data. We note that the exact size for the temporal dimension is not crucial, since we will also leverage random cropping and resizing along the temporal dimension for data augmentation. Since these tactile images only have a single channel, the channel is repeated three times to match the input dimension for the vision foundational model used in the experiments. All tactile images and GelSight data is normalized to the range of $[-1, 1]$.

Model Architecture and Pre-training. We choose a ResNet-18 pre-trained on MetaDataset [45] as our foundational vision model. It is chosen for its balanced accuracy and computational efficiency. We emphasize that other foundational models may be easily chosen given the trade-off between accuracy and efficiency. We also highlight all experiments use the *identical* foundational model without any modification, as our approach allows fully shared models.

A. Standard Material Classification

We compare our approach with baseline methods on standard material classification using BioTac-20, RoboSkin and GelSight datasets. We highlight that the baselines are specifically tailored to the BioTac or RoboSkin sensors, whilst our model is generic.

TABLE III: Material Classification Accuracy (%). Numbers for baseline methods are originally reported in [8]. Pre-train denotes initialization with the foundational vision model.

Method	BioTac-20	RoboSkin	GelSight
SVM	94.2 ± 0.7	50.5 ± 5.6	n.a
SVM (spikes)	93.5 ± 1.5	63.3 ± 1.8	n.a
Conv-LSTM	94.5 ± 1.5	93.5 ± 0.5	n.a
SNN	94.6 ± 1.3	92.2 ± 0.5	n.a
Least Square w/ Pre-train	93.8 ± 1.2	84.8 ± 1.3	67.1 ± 0.8
ResNet (ours)	98.0 ± 0.3	95.0 ± 0.6	92.9 ± 0.3
ResNet w/ Pre-train (ours)	98.9 ± 0.2	96.0 ± 0.5	95.1 ± 0.3

Model Details. Our model is trained for 100 epochs using stochastic gradient descent (SGD). A validation set is employed to schedule the learning rate, mitigating performance plateaus. An initial learning rate of 0.01 is chosen empirically, with a momentum of 0.9 and a weight decay of 0.0001. 5-fold cross validation is performed for all experiments.

Baseline Methods. We compare our approach to a diverse set of methods investigated in [8], including a spiking neural network (SNN), LSTM, regular support vector machine (SVM) and spike-encoded SVM (SVM Spike).

Table III reports the classification accuracy for all evaluated methods. Our generic ResNet outperforms the baselines by more than 4%, suggesting the viability of our tactile image approach. In addition, the results clearly shows that fine-tuning from the foundational model is more advantageous than random initialization. This indicates positive knowledge transfer from the pre-trained model and improved generalization. This is especially visible for the GelSight dataset owing to the imbalance between the small size of the dataset and the large input dimension.

Pre-training also noticeably improves learning efficiency, as reported in Fig. 4. For both BioTac-20 and RoboSkin datasets, transferring from the foundational model (i.e., with pre-training) achieves higher accuracy with fewer iterations over the training data. Learning efficiency is a desirable property for robots requiring fast adaptation to novel environments.

Foundational Models and Tactile Images. To better understand the connection between our foundational model and tactile images, we introduce another baseline in Tab. III denoted by ‘‘Least Square’’. This baseline encodes all tactile images into fixed representations using the pre-trained ResNet, and only learns a least-squares classifier over the fixed representation. The accuracy of this baseline thus directly reflects the usefulness of the pre-trained model towards tactile images. Surprisingly, the results show that the foundational vision model trained from natural images already contains the general features required for tactile texture representation, despite the apparent distributional shift. This provides direct support to the connection between visual and tactile perception, resembling the similarities between the human visual cortex and somatosensory system. The results also provide empirical justification for our choice of tactile images as model input.

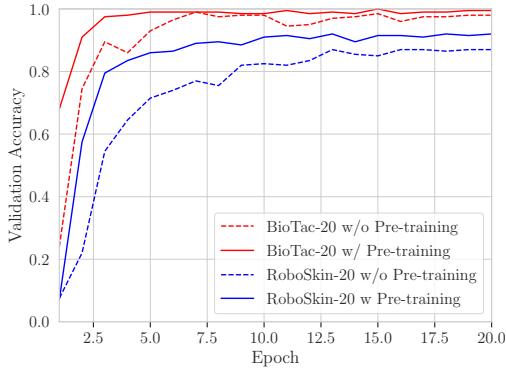


Fig. 4: Test Accuracy over the first 20 epochs for both BioTac-20 (red) and RoboSkin-20 (blue), with (solid) and without (dashed) pre-training.

B. Augmentation and Model Robustness

As noted in Sec. IV-B, data augmentations applied to tactile images may be interpreted as diversifying the conditions of data collection. This is crucial for tactile datasets as they are generally expensive to collect. We investigate the effects of augmentation in the following experiments.

Robustness to Sampling Length. For material classification, it is desirable to shorten the sampling length without sacrifice to accuracy. This corresponds to classifying randomly cropped tactile images in our formulation. It was also investigated in [8] as a strength of spiking neural architecture. In Fig. 5, we investigate how random cropping affects classification accuracy over varying data length, and compare our approach to previous methods.

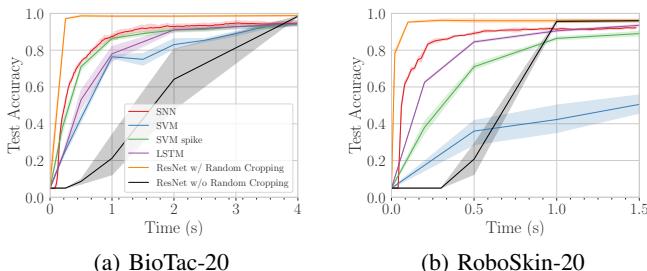


Fig. 5: Test accuracies of our approach with and without random cropping augmentations for varying data length. Baseline methods included for comparison.

The results clearly show that our model with augmentation outperforms the previous methods, achieving higher test accuracy with less data required. For both datasets, ResNet with augmentation is able to accurately classify the materials with about 0.3 seconds of sensor data. As the data length increases, the test accuracy rapidly increases and remains high, suggesting that our model could efficiently accumulate information over short duration while maintaining robustness over long run. In addition, Fig. 5 shows that augmentation

is crucial for robust performance. The same model trained without augmentation performed the worst among all methods, suggesting overfitting to the original data length and less robust features learned.

Robustness to Movement Speed. While some tactile datasets are collected under a tightly controlled robot motion, it is preferable that the learned model generalizes to more varied motions. We simulate different speeds of the robot's sliding motion during tactile sensing by sub-sampling the test set data along the temporal axis, and investigate the effects of augmentation on this out-of-distribution test set.

Fig. 6 shows that the model trained with random resizing augmentation is robust against varying robot speed, achieving consistent accuracy across different movement speed. In contrast, the model with no augmentation generalized poorly even with slight speed deviation. The figure also shows that random cropping improves model robustness against varying movement speed.

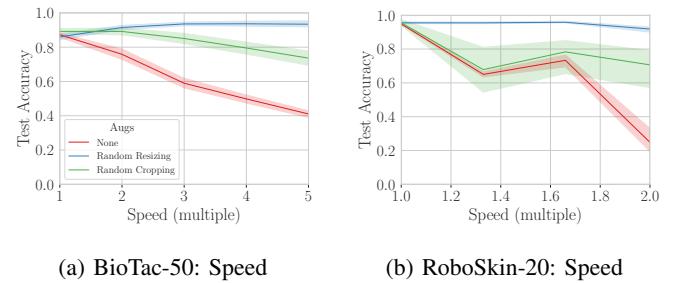


Fig. 6: The effects of augmentation with respect to varying robot movement speed during tactile sensing. X-axis denotes the multiples of the original robot speed.

Robustness to Sensor Noise. Similar to the previous experiment, we construct another out-of-distribution test set by injecting random sensor noise. Fig. 7 and evaluates the effects of augmentations.

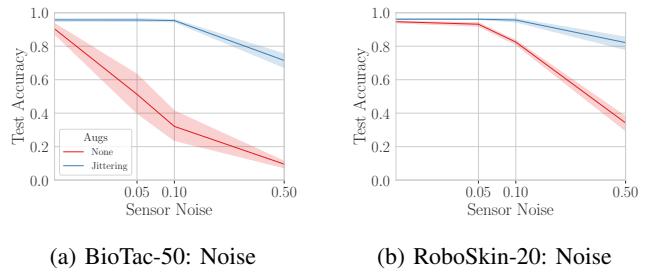


Fig. 7: The effect on test accuracy with respect to sensor noise. X-axis denotes maximum noise level added to tactile images.

Fig. 7 shows that model trained without random jittering augmentation generalizes poorly to noisy data, especially on BioTac dataset. This is due to the BioTac data being collected under a strict condition, including fixed force and movement speed. The model trained on non-augmented BioTac data thus

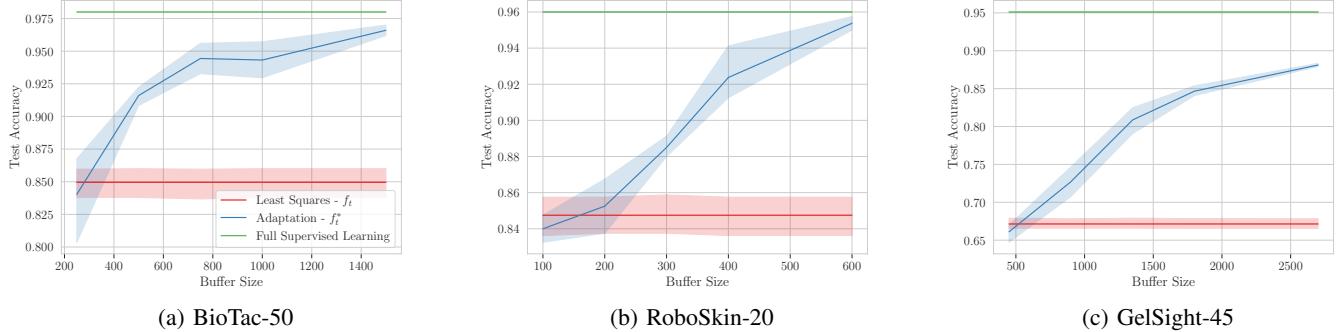


Fig. 8: CL performance across all BioTac-50, RoboSkin-20, and GelSight-45 datasets, for varying buffer sizes. Accuracy from supervised upper bound and ridge regression are shown to illustrate the performance changes associated with adaptation. With increasing memory buffer, CL achieves better test accuracy and narrows the gap against standard supervised learning.

overfits to the homogeneous data and lack robustness. In contrast, RoboSkin data contains more diverse samples since it is collected without strict speed or force control. As reflected in Fig. 7, the non-augmented model trained on RoboSkin data is therefore naturally robust to a low level of sensor noise. However, as the noise level increases, the test accuracy of all non-augmented models still deteriorate rapidly.

Fig. 7 also indicates that the model trained with augmentation can significant sensor noise, with the noise level of 0.5 representing a potentially 50% deviation from the intended value range. At this level, the augmented model still retains a test accuracy of 80% for RoboSkin and 73% for BioTac-50. Lastly, we observe that even for the original test set (i.e., noise level = 0), the augmented model still outperforms the non-augmented version, suggesting more robust features learned with augmentation.

Overall, we have demonstrated that standard CV augmentations can be directly applied to tactile images to appreciably boost model robustness in various aspects, including sampling length, movement speed and sensor noise. As several of our experiments relied on simulated test data, we will further demonstrate the usefulness of augmentation with real out-of-distribution data in Sec. V-D.

C. Continual Tactile Representation Learning

As described in Sec. II-A, we cast material classification in a CL setting, which requires our model to learn each material sequentially. CL enables robots to continuously acquire new tactile experiences, without having to perform expensive retraining from scratch.

Model Detail. The same foundational vision model is used as the embedding model for Alg. 1. During fine-tuning with memory buffer M_t , we adopt data augmentation and a cosine learning schedule [46] to mitigate overfitting. For all experiments, we perform a 5-fold cross-validation.

Fig. 8 shows the CL performance for each dataset over different memory buffer sizes. We report the performance of f_t and the fine-tuned f_t^* . We also include the test accuracy of standard material classification as a performance reference.

Note that f_t obtained via recursive least squares is equivalent to the least-squares baseline discussed in Sec. V-A. Thanks to the foundational vision model, f_t thus guarantees a robust minimum performance level for CL (see red lines in Fig. 8). f_t^* is obtained by adapting f_t with the memory buffer. Its performance improves with larger memory buffers, closing the gap with standard material classification. For BioTac and RoboSkin particularly, the CL performance is comparable with standard supervised learning, using a moderate memory buffer of 1500 and 600 respectively. The memory buffer required only represents a fraction of the original datasets, suggesting that our approach also allows efficient and accurate CL of new materials with limited memory requirements.

D. Fabric Composition Detection

Introduced in Sec. II-A, fabric composition detection involves predicting the presence of six constituent materials, including Linen, Viscose, Cotton, Wool, Polyester and Elastane, in different fabrics. A single model is learned to detect the presence of all constituents concurrently, with one prediction head for each constituent. This task is more challenging than standard material classification, due to the “similar feels” of different fabrics. The physical weave of a fabric also contributes to its feel, adding a potential confounding factor for the task.

For this task, the data is collected using Contactile sensor. As discussed in Sec. III, we deliberately used two protocols for data collection. The training set is collected using strict force and velocity control while the test set is collected with more natural movements. The test set thus presents a more realistic setting and a clear domain shift with respect to the training data.

Model Details. The training procedure is similar to that used for standard material classification. The only change is that the number of training epochs is reduced from 100 to 50. Data augmentations are applied to model training when specified. For evaluation, we consider the average classification score for all constituents materials. For instance, Felt contains Viscose and Wool. The learned model only achieves a score of 1 for

predicting precisely the two constituents. Any false positive or false negative detection will decrease the score by $\frac{1}{6}$.

Tab. IV shows the average classification score for different model setups. We investigate both knowledge transfer from foundational vision model and model pre-trained on other sensors. We also study the effects of data augmentation.

TABLE IV: Fabric Composition Detection Accuracy (%)

Model	Test Accuracy Score
Least Squares w/ vision Pre-train	74.2
Least Squares w/ BioTac Pre-Train	76.1
ResNet	76.3
ResNet + Augmentation	78.9
ResNet + Augmentation (BioTac Pre-train)	80.6

In Tab. IV, we again leverages least-squares classifier over a fixed representation to quantify the effectiveness of a pre-trained model. We see that directly applying the foundational vision model achieves 74.2%, while applying the BioTac model obtained in Sec. V-A achieves 76.1%. The result is our first demonstration of *successful cross-task and cross-sensor transfer*: the BioTac model trained on standard material classification can be directly applied to Contactile data for fabric composition detection. This result demonstrates the general applicability of our approach, and its ability for robust and flexible knowledge transfer.

Tab. IV further demonstrates the usefulness of data augmentations on real out-of-distribution data, with augmentation contributes over 2% in test accuracy compared to the non-augmented model. The results validate our physical interpretations for the applied augmentations, showing that the augmented model is indeed more robust against more varied motions. From another perspective, we may also leverage the synthetic data produced by augmentation to reduce data collection load. This is important if a robot is only allowed limited (exploratory) interaction with environments. Lastly, we remark that the best model is obtained by combining both knowledge transfer and augmentation, achieving 80.6% in test accuracy.

E. Observations on the Learned Representation

Results from previous sections suggest robust knowledge transfer across sensors despite the varied sensing mechanisms and data format. We hypothesize that this could be a result of a learned invariant descriptor of the tactile properties of the contact surfaces. Since the processing of texture in the human somatosensory cortex is a relatively lower-level function, we are thus interested in understanding if the lower-level abstraction in the learned model recovers similar latent representation for diverse sensor data.

Fig. 9 shows the feature activation for different sensors using Deep Dream technique [47]. This qualitative visualization of the learned features shows that feature activation generated right after the first block for 3 ResNets, each fine-tuned on a separate tactile dataset in standard material classification.

All 3 feature activation maps have high resemblance of one another, suggesting that learned model indeed recovers consistent representation of tactile properties despite diverse sensing mechanisms. This further supports the knowledge transferrability between different sensors and related tasks.

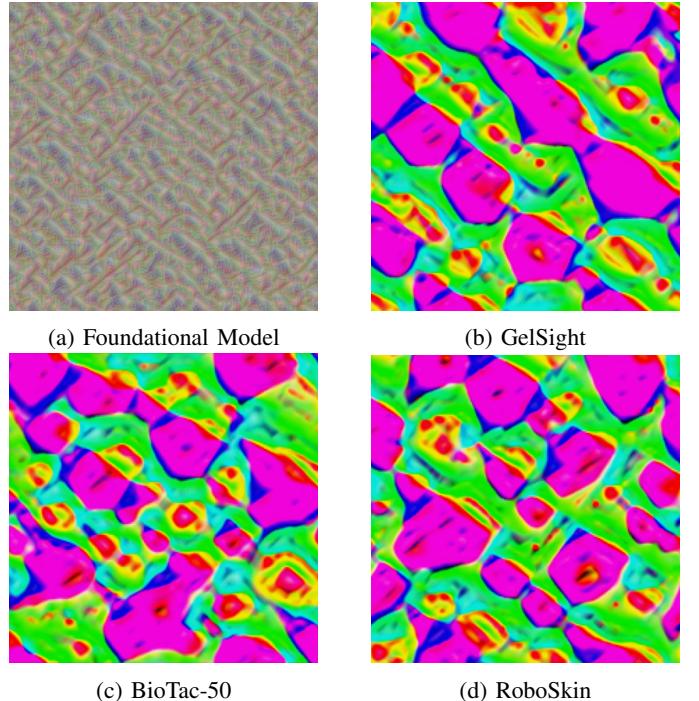


Fig. 9: Feature activation after block 1 of ResNet. (a) Original feature activation from foundational CV model. (b), (c), (d) Feature activation after fine-tuning with specific sensor data.

VI. CONCLUSION

In this work, we presented a foundational model approach to tactile representation learning. In contrast to sensor-specific tactile models, our approach is characterized by a standardized ML pipeline, including a unifying data format for diverse tactile data, fully shared model architecture and learning techniques, all of which are key requirements for foundational models. Further, the experiment results suggest that our approach not only outperforms sensor-specific models, but crucially allows efficient knowledge transfer between models trained on different sensors and tasks, satisfying the remaining property for foundational models. In particular, we demonstrated the connection between visual and tactile perception, showing that foundational vision models trained on natural images can be a readily accessible source of knowledge for tactile representation learning. This also allows us to effectively perform, with the same unified model, downstream tasks which were previously achieved with an array of methods in the literature. We believe that this investigation thus contributes a robust and general approach to tactile representation learning and provides a strong baseline for future research.

REFERENCES

- [1] Huaping Liu, Yupei Wu, Fuchun Sun, and Di Guo. Recent progress on tactile object recognition. *International Journal of Advanced Robotic Systems*, 14(4):1729881417717056, 2017. doi: 10.1177/1729881417717056. URL <https://doi.org/10.1177/1729881417717056>.
- [2] Shan Luo, Xiaozhou Liu, Kaspar Althoefer, and Hongbin Liu. Tactile object recognition with semi-supervised learning. In *Intelligent Robotics and Applications: 8th International Conference, ICIRA 2015, Portsmouth, UK, August 24-27, 2015, Proceedings, Part II 8*, pages 15–26. Springer, 2015.
- [3] Marco Costanzo, Giuseppe De Maria, and Ciro Natale. Two-fingered in-hand object handling based on force/tactile feedback. *IEEE Transactions on Robotics*, 36(1):157–173, 2020. doi: 10.1109/TRO.2019.2944130.
- [4] Osher Azulay, Inbar Ben-David, and Avishai Sintov. Learning haptic-based object pose estimation for in-hand manipulation with underactuated robotic hands, 2022. URL <https://arxiv.org/abs/2207.02843>.
- [5] Wenzhen Yuan, Siyuan Dong, and Edward H. Adelson. Gelsight: High-resolution robot tactile sensors for estimating geometry and force. *Sensors*, 17(12), 2017. ISSN 1424-8220. doi: 10.3390/s17122762. URL <https://www.mdpi.com/1424-8220/17/12/2762>.
- [6] Jeremy A Fishel and Gerald E Loeb. Sensing tactile microvibrations with the biotac—comparison with human sensitivity. In *2012 4th IEEE RAS & EMBS international conference on biomedical robotics and biomechatronics (BioRob)*, pages 1122–1127. IEEE, 2012.
- [7] Heba Khamis, Raquel Izquierdo Albero, Matteo Salerno, Ahmad Shah Idil, Andrew Loizou, and Stephen J Redmond. Papillarray: An incipient slip sensor for dexterous robotic or prosthetic manipulation—design and prototype validation. *Sensors and Actuators A: Physical*, 270:195–204, 2018.
- [8] Tasbolat Taunyazov, Yansong Chua, Ruihan Gao, Harold Soh, and Yan 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, 2020. doi: 10.1109/IROS45743.2020.9340693.
- [9] Anupam Kumar Gupta, Andrei Nakagawa-Silva, Nathan F. Lepora, and Nitish V. Thakor. Spatio-temporal encoding improves neuromorphic tactile texture classification. *IEEE Sensors Journal*, 21(17):19038–19046, 2021. doi: 10.1109/JSEN.2021.3087511.
- [10] Ruihan Gao, Tasbolat Taunyazov, Zhiping Lin, and Yan Wu. Supervised autoencoder joint learning on heterogeneous tactile sensory data: Improving material classification performance. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10907–10913, 2020. doi: 10.1109/IROS45743.2020.9341111.
- [11] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.
- [12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.
- [13] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.
- [14] Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Self-supervised learning of local visual features. *arXiv preprint arXiv:2210.01571*, 2022.
- [15] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- [16] Steven S. Hsiao and Manuel Gomez-Ramirez. *Neural Mechanisms of Tactile Perception*, chapter 8. 2012. ISBN 9781118133880. doi: <https://doi.org/10.1002/9781118133880.hop203008>. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118133880.hop203008>.
- [17] Michal Haindl and Jiří Filip. Visual texture: Accurate material appearance measurement, representation and modeling. 2013.
- [18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.
- [19] Jeremy A Fishel and Gerald E Loeb. Bayesian exploration for intelligent identification of textures. *Frontiers in neurorobotics*, 6:4, 2012.
- [20] Janine Hoelscher, Jan Peters, and Tucker Hermans. Evaluation of tactile feature extraction for interactive object recognition. In *2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids)*, pages 310–317. IEEE, 2015.
- [21] Tasbolat Taunyazov, Weicong Sng, Hian Hian See, Brian Lim, Jethro Kuan, Abdul Fatir Ansari, Benjamin Tee, and Harold Soh. Event-driven visual-tactile sensing and learning for robots. In *Proceedings of Robotics: Science and Systems*, July 2020.
- [22] Shiv S Baishya and Berthold Bäuml. Robust material classification with a tactile skin using deep learning. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8–15. IEEE, 2016.
- [23] Ruihan Gao, Tian Tian, Zhiping Lin, and Yan Wu. On explainability and sensor-adaptability of a robot tactile texture representation using a two-stage recurrent networks. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1296–1303,

2021. doi: 10.1109/IROS51168.2021.9636380.
- [24] Nawid Jamali and Claude Sammut. Majority voting: Material classification by tactile sensing using surface texture. *IEEE Transactions on Robotics*, 27(3):508–521, 2011. doi: 10.1109/TRO.2011.2127110.
- [25] W.W. Mayol-Cuevas, J. Juarez-Guerrero, and S. Munoz-Gutierrez. A first approach to tactile texture recognition. In *SMC'98 Conference Proceedings. 1998 IEEE International Conference on Systems, Man, and Cybernetics (Cat. No.98CH36218)*, volume 5, pages 4246–4250 vol.5, 1998. doi: 10.1109/ICSMC.1998.727512.
- [26] Nicholas Pestell and Nathan Lepora. Artificial sa-i, ra-i and ra-ii/vibrotactile afferents for tactile sensing of texture. *Journal of The Royal Society Interface*, 19, 04 2022. doi: 10.1098/rsif.2021.0603.
- [27] Harold Soh, Yanyu Su, and Yiannis Demiris. Online spatio-temporal gaussian process experts with application to tactile classification. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4489–4496. IEEE, 2012.
- [28] Harold Soh and Yiannis Demiris. Incrementally learning objects by touch: Online discriminative and generative models for tactile-based recognition. *IEEE transactions on haptics*, 7(4):512–525, 2014.
- [29] Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3366–3385, 2021.
- [30] Subramanian Sundaram, Petr Kellnhofer, Yunzhu Li, Jun-Yan Zhu, Antonio Torralba, and Wojciech Matusik. Learning the signatures of the human grasp using a scalable tactile glove. *Nature*, 569:698–702, 05 2019. doi: 10.1038/s41586-019-1234-z.
- [31] Jivko Sinapov, Vladimir Sukhoy, Ritika Sahai, and Alexander Stoytchev. Vibrotactile recognition and categorization of surfaces by a humanoid robot. *IEEE Transactions on Robotics*, 27(3):488–497, 2011. doi: 10.1109/TRO.2011.2127130.
- [32] Alexander Schmitz, Marco Maggiali, Lorenzo Natale, Bruno Bonino, and Giorgio Metta. A tactile sensor for the fingertips of the humanoid robot icub. In *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2212–2217. IEEE, 2010.
- [33] Benjamin Ward-Cherrier, Nicholas Pestell, Luke Cramphorn, Benjamin Winstone, Maria Elena Giannaccini, Jonathan M. Rossiter, and Nathan F. Lepora. The tactip family: Soft optical tactile sensors with 3d-printed biomimetic morphologies. *Soft Robotics*, 5:216 – 227, 2018.
- [34] Danfei Xu, Gerald E. Loeb, and Jeremy A. Fishel. Tactile identification of objects using bayesian exploration. In *2013 IEEE International Conference on Robotics and Automation*, pages 3056–3061, 2013. doi: 10.1109/ICRA.2013.6631001.
- [35] Nicholas Wettels, Veronica J. Santos, Roland S. Johansson, and Gerald E. Loeb. Biomimetic tactile sensor array. *Advanced Robotics*, 22(8):829–849, 2008. doi: 10.1163/156855308X314533. URL <https://doi.org/10.1163/156855308X314533>.
- [36] Tasbolat Taunyazov, Hui Fang Koh, Yan Wu, Caixia Cai, and Harold Soh. Towards effective tactile identification of textures using a hybrid touch approach. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 4269–4275, 2019. doi: 10.1109/ICRA.2019.8793967.
- [37] Zhanat Kappassov, Juan-Antonio Corrales, and Véronique Perdereau. Tactile sensing in dexterous robot hands. *Robotics and Autonomous Systems*, 74: 195–220, 2015.
- [38] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
- [39] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.
- [40] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier, 1989.
- [41] Roger Ratcliff. Connectionist models of recognition memory: constraints imposed by learning and forgetting functions. *Psychological review*, 97(2):285, 1990.
- [42] Ruohan Wang, Marco Ciccone, Giulia Luise, Massimiliano Pontil, Andrew Yapp, and Carlo Ciliberto. Schedule-robust online continual learning. *arXiv preprint arXiv:2210.05561*, 2022.
- [43] Thomas Kailath, Ali H Sayed, and Babak Hassibi. *Linear Estimation*. Prentice Hall, 2000.
- [44] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2001–2010, 2017.
- [45] Eleni Triantafillou, Tyler Zhu, Vincent Dumoulin, Pascal Lamblin, Utku Evci, Kelvin Xu, Ross Goroshin, Carles Gelada, Kevin Swersky, Pierre-Antoine Manzagol, and Hugo Larochelle. Meta-dataset: A dataset of datasets for learning to learn from few examples. In *International Conference on Learning Representations*, 2020. URL <https://openreview.net/forum?id=rkgAGAVKPr>.
- [46] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.
- [47] A. Mordvintsev, Christopher Olah, and Mike Tyka. Inceptionism: Going deeper into neural networks. 2015.