Lifelong Visual-Tactile Cross-Modal Learning for Robotic Material Perception

Wendong Zheng, Huaping Liu, *Senior Member, IEEE*, and Fuchun Sun, *Fellow, IEEE*

***Abstract*— The material attribute of an object's surface is critical to enable robots to perform dexterous manipulations or actively interact with their surrounding objects. Tactile sensing has shown great advantages in capturing material properties of an object's surface. However, the conventional classification method based on tactile information may not be suitable to estimate or infer material properties, particularly during interacting with unfamiliar objects in unstructured environments. Moreover, it is difficult to intuitively obtain material properties from tactile data as the tactile signals about material properties are typically dynamic time sequences. In this article, a visual-tactile cross-modal learning framework is proposed for robotic material perception. In particular, we address visual-tactile cross-modal learning in the lifelong learning setting, which is beneficial to incrementally improve the ability of robotic cross-modal material perception. To this end, we proposed a novel lifelong cross-modal learning model. Experimental results on the three publicly available data sets demonstrate the effectiveness of the proposed method.**

***Index Terms*— Cross-modal learning, lifelong learning, material recognition, robotic perception.**

I. Introduction

IN TASKS of dexterous manipulation or actively interacting with its objects, intelligent robots need to understand the material properties of the target's surface before performing some certain actions [1]. According to these properties of information, the robot can select an appropriate manipulation strategy or interaction pattern [2]–[4]. In [5]–[8], several promising techniques, such as operation space transformation, decoupled force/motion control, variable stiffness technique, and human-compliant cooperation/coordination, have proposed to improve the performance of autonomous system. These works are significant in human-centered robotics scenario and represent the state-of-the-art research in such robots. Surface material perception plays a critical role in many fields involving robotic manipulation and environmental interaction [9]. Recently, research about robotic material recognition has received increasing attention.

For perceiving material properties in the robotic field, tactile sensing has shown great advances in capturing information about material properties [10]. With the development of tactile sensors, tactile modality has been widely used for material recognition [11]. However, the conventional classification method based on tactile information may not be suitable to estimate or infer material properties, particularly during interacting with unfamiliar objects in unstructured environments [12], [13]. Moreover, it is difficult to intuitively obtain material properties from tactile data as these tactile signals delivering material properties are typically dynamic high-dimensional time series.

It is well known that humans can quickly distinguish or estimate material tactile properties of an unfamiliar object surface simply by visual images. It is mainly due to that human can transfer knowledge about material properties from tactile modality to visual modality by building cross-modal correlations [14]. Baumgartner *et al.* [15] and Fleming [16] also demonstrated that material tactile properties and visual perception are highly correlated for human by performing the experiments. The mechanism of cross-modal material perception is hopeful to help robot perform various manipulation or interaction tasks in practical applications more effectively, particularly in remote unstructured scenarios.

Motivated by the mechanism, we sought to enable robot to achieve the ability of cross-modal material perception in some operation tasks. Specifically, our goal is to enable robots to explore objects autonomously using tactile sensors and report a set of surface images that describe material properties. By this cross-modal retrieval way, material properties of unknown objects surfaces can be judged or estimated, which is useful for robotic manipulation or interaction in remote unstructured environments [8], [17]. The core of this problem is visual-tactile cross-modal learning.

Currently, robotic learning typically runs a machine-learning algorithm on a given data set to generate a model, and then, the learned model is applied in the specified task [18]. In this classic isolated learning paradigm, the machine-learning algorithms typically require large numbers of training data to learn an effective model [19]. However, collecting large amounts of data is often unfeasible in many robotic applications [20]. In particular, collecting large-scale tactile data is extremely

Manuscript received September 19, 2019; revised January 8, 2020; accepted March 3, 2020. This work was supported in part by the National Key Research and Development Program under Grant 2018YFB1305102. This work was completed while Wendong Zheng was visiting Tsinghua University, Beijing, China. *(Corresponding author: Huaping Liu.)*

Wendong Zheng is with the State Key Laboratory of Reliability and Intelligence of Electrical Equipment, Hebei University of Technology, Tianjin 300130, China, and also with the Key Laboratory of Electromagnetic Field and Electrical Apparatus Reliability of Hebei Province, School of Electrical Engineering, Hebei University of Technology, Tianjin 300130, China (e-mail: zwendong@126.com).

Huaping Liu and Fuchun Sun are with the Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China, and also with the State Key Laboratory of Intelligent Technology and Systems, Beijing National Research Center for Information Science and Technology, Beijing 100084, China (e-mail: hpliu@tsinghua.edu.cn; fcsun@tsinghua.edu.cn).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2020.2980892

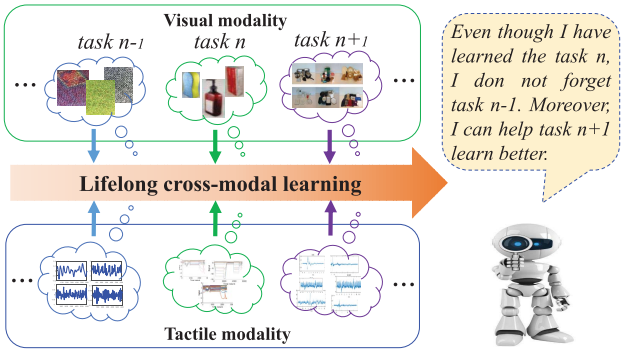


Fig. 1. Schematic of lifelong visual-tactile cross-modal learning. The same color boxes denote that visual and tactile data come from the same task.

difficult due to the dynamic collection process and diversity of tactile modality.

On the other hand, lifelong learning [21] is proposed to imitate the human learning process and capability, which aims to overcome the limitation of the isolated learning paradigm. Its goal is to incrementally learn a sequence of tasks, retain the knowledge learned from previous tasks, and use the accumulated knowledge both to help learn new tasks and to improve the performance of previous tasks [22], [23]. In other words, the model with its prior knowledge can be learned well on a relevant new task, even using the small amount of training data from a new task. These merits of the lifelong learning motivate us to apply it in the problem of visual-tactile cross-modal learning. To the best of our knowledge, none of the existing work performs visual-tactile cross-modal learning in the lifelong learning context.

In this article, we propose to address visual-tactile cross-modal learning in the lifelong learning setting. A brief illustration of lifelong visual-tactile cross-modal learning is shown in Fig. 1, where robots need to incrementally learn a sequence of tasks. At the same time, the ability of robotic cross-modal learning is expected to become increasingly powerful as it learns more and more tasks. This article exhibits the twofold challenges, which are cross-modal learning of heterogeneous modalities and continual learning of multiple tasks. To this end, we propose a novel lifelong cross-modal learning (LLCL) method, which consists of intratask cross-modal correlation learning subnetwork and intertask correlation lifelong learning subnetwork. The main contributions are briefly summarized as follows.

1) Visual-tactile cross-modal learning in the lifelong learning setting is proposed. The framework aims to reuse the previous learning knowledge to help learn a new task. Finally, the learned model not only works well for the new task but also preserves or even enhance performance on the previous task.
2) A novel lifelong visual-tactile cross-modal learning method is developed, which incorporates advantages of lifelong learning and deep adversarial learning to explore pairwise correlation of heterogeneous data from a sequence of tasks.
3) We perform experimental validation on the publicly available data set. The experimental results demonstrate the validity of the proposed intratask cross-modal correlation learning subnetwork in each single task. Moreover, the proposed LLCL achieves better performance than individual single-task learning. It indicates that the proposed LLCL model not only helps correlation learning of new tasks but also boosts the correlation learning ability on previous tasks.

Some methods have been proposed to solve the problem of visual-tactile cross-modal learning [11], [24]–[31]. The works of [24], [25], [27], and [28] still focused on material recognition, where the associating tactile information and visual information is used as complementary information to improve the performance of either vision or tactile sensing. In [26] and [29], they associated tactile information and visual information to train an embedding space, where an input image can be encoded into the correspondence tactile information. The work [31] focused on cross-modal prediction of object instance. Although the idea of visual-tactile cross-modal retrieval for material perception was studied in our previous works [11] and [30], they did not explicitly consider group-invariant characteristic of weakly paired matching. Moreover, most of them only focused on single-task learning and adopted a paradigm of off-line training and online performing. The research on lifelong learning is still vacant in visual-tactile cross-modal learning. Obviously, this article is different from these works in terms of research objective and application scenario.

In the following, Section II briefly reviews some related works. In Section III, we formulate the problem of the visual-tactile cross-modal lifelong learning for robotic material perception. Section IV details the network architecture and learning algorithm of the proposed method. Experimental results and analysis are represented in section V.

II. Related Work

In this section, some related works are briefly reviewed from two aspects: cross-modal retrieval and lifelong learning.

A. Cross-Modal Retrieval

Cross-modal retrieval aims to take data of one modality as the query to search relevant data of another modalities [32]. Its main challenge is that their similarity of data from different modalities cannot be calculated directly, which is due to the differences of the representations and distributions between different modalities. An intuitive solution is to project pairwise data of different modalities into a common space, where their content similarity can be measured by common distance metrics. According to this idea, many methods [33]–[36] have been proposed, which aims to learn a pair of linear projection matrices for common representation of heterogeneous modalities. Canonical correlation analysis (CCA) [33] is the representative of these methods, which maximizes the correlation of pairwise samples of different modalities to learn projection matrices. Although these methods have been applied in some tasks involving cross-modal learning, they may not be able to effectively model complex correlation by projection matrices.

Since deep networks can effectively model complex transformations, it has been introduced into cross-modal correlation learning of heterogeneous data, such as [37]–[40]. Although these methods achieved promising performance in some cross-modal retrieval tasks, their optimization objectives focus only on pairwise relevance [41], [42]. Recently, metric learning [43] and adversarial learning [44] have been incorporated into common representation learning, such as DAML [45] and ACMR [46], which show promising performance in tasks of image-text retrieval.

However, these methods may not be suitable for visual-tactile cross-modal learning tasks. It is mainly due to that these methods do not consider large intraclass variance in their models, which demonstrates prominence in weakly paired visual-tactile cross-modal learning tasks. Moreover, almost all of the existing methods of cross-modal learning cannot be directly used for continual learning, where training data of different tasks become incrementally available over time. The ability of continuous learning over time by acquiring new knowledge from new tasks while retaining experiences learned from the previous tasks is crucial for autonomous agents and robots interacting with the environment [47], [48].

B. Lifelong Learning

The current dominant machine-learning algorithms almost focus on data-driven optimization learning [49]. Namely, it is to run a machine-learning algorithm on a given training data to learn a model. Then, the learned model is used in its specified application. This paradigm does not consider to retain and accumulate the previously learned knowledge and use it in feature learning [50], [51]. Unlike this isolated learning paradigm, human can adapt our learned knowledge to seamlessly solve new problems and also learn from it [52], [53]. Therefore, we can learn more and more knowledgeable over time and become more effective for problem-solving. Without the ability of accumulating and reusing the previously learned knowledge, a machine-learning algorithm typically needs sufficient training data in order to learn an effective model. In fact, collecting a large number of data is difficult or even impossible in many application scenarios.

Lifelong learning is proposed to try to imitate the continuous learning process of human, which aims to overcome the drawbacks of the isolated data-driven optimization learning. Its objective is to incrementally learn a sequence of tasks, retain the knowledge from previous tasks, and exploit the accumulated knowledge to improve future learning [22], [54]. Recently, lifelong learning gains increasing attention in deep learning context due to the promising performance of deep neural networks. However, a learned model using deep neural networks usually suffers from catastrophic forgetting when it adapts to a new task [55], [56]. Specifically, when training on a new task, the learned model tends to forget the knowledge learned from previous tasks. This means that a new task will override the network parameters learned in the past, thus abruptly degrading the performance of the learned model in the previous tasks. Intuitively, retraining the model with both training data of previous tasks and training data of new task can mitigate the effects of catastrophic interference [57]. However, it often relies on large storage system to store the past data, which is typically infeasible in real-world robotic applications. Recently, many methods [58]–[62] have been proposed for lifelong learning. However, the existing works mainly focus on classification tasks within single modality, which cannot directly handle sequences tasks of visual-tactile cross-modal learning.

III. Problem Formulation

In this article, the problem we mainly solve is visual-tactile cross-modal learning in the lifelong learning setting. The goal is to incrementally learn the pairwise correlation of heterogeneous visual and tactile data from a sequence of tasks. Namely, when a new task arrives, we aim to reuse and transfer the knowledge gained from the previous tasks to help learn cross-modal correlation in the new task and further improve the correlation learning ability of the model on the previously learned tasks.

A. Problem Description

Let $S = \{S_1, S_2, \ldots, S_j\}$ be a collection of j tasks for lifelong visual-tactile cross-modal learning. We denote the task S_j as $D_{\text{tr}} = \{a_j^i\}_{i=1}^N$, $a_j^i = \{(u_j^i, v_j^i), y_j^i\}$, where (u_j^i, v_j^i) is ith sample pair of the visual feature $u_j^i \in R^{d_u}$ and the tactile feature $v_j^i \in R^{d_v}$ with label y_j^i. d_u and d_v are the corresponding dimension of visual and tactile feature. Similarly, the task S_{j+1} is denoted as $\{(u_{j+1}^i, v_{j+1}^i), y_{j+1}^i\}$. The tasks S_j and S_{j+1} have different instance spaces, i.e., $S_j \cap S_{j+1} = 0$.

As shown in Fig. 1, visual and tactile are two completely different sensing modalities for robots, which entails that they have different feature representations. For instance, visual data are static images from the space of pixels, while tactile are dynamic time series from the space of time domain. Therefore, their content similarity cannot be directly measured by a common distance metric. In fact, modality discrepancy is the fundamental challenge of cross-modal learning tasks, such as text-images retrieval. To relieve this issue, many methods [37]–[40] are proposed for cross-modal learning, most of which mainly focus on maximizing paired samples correlation of heterogeneous modalities. However, they usually ignore the distribution discrepancy of feature representation, so the learned common representations are still highly heterogeneous. It limits the performance of cross-modal learning.

In addition, visual data and tactile data are typically collected separately in robotic applications. Moreover, visual data are obtained by a camera from large surface areas, while tactile data are dynamically collected by a tactile sensor from a small contact region. Therefore, there is no one-to-one correspondence between the visual sample and the tactile sample. Instead, a group of samples from tactile modality is matched to a group of samples in the visual modality based on category information, which is known as weakly paired matching [30]. Unfortunately, most of the existing methods of cross-modal learning consider one-to-one paired data to learn an effective model for common representation, which may not be suitable for weakly paired visual-tactile cross-modal learning tasks.

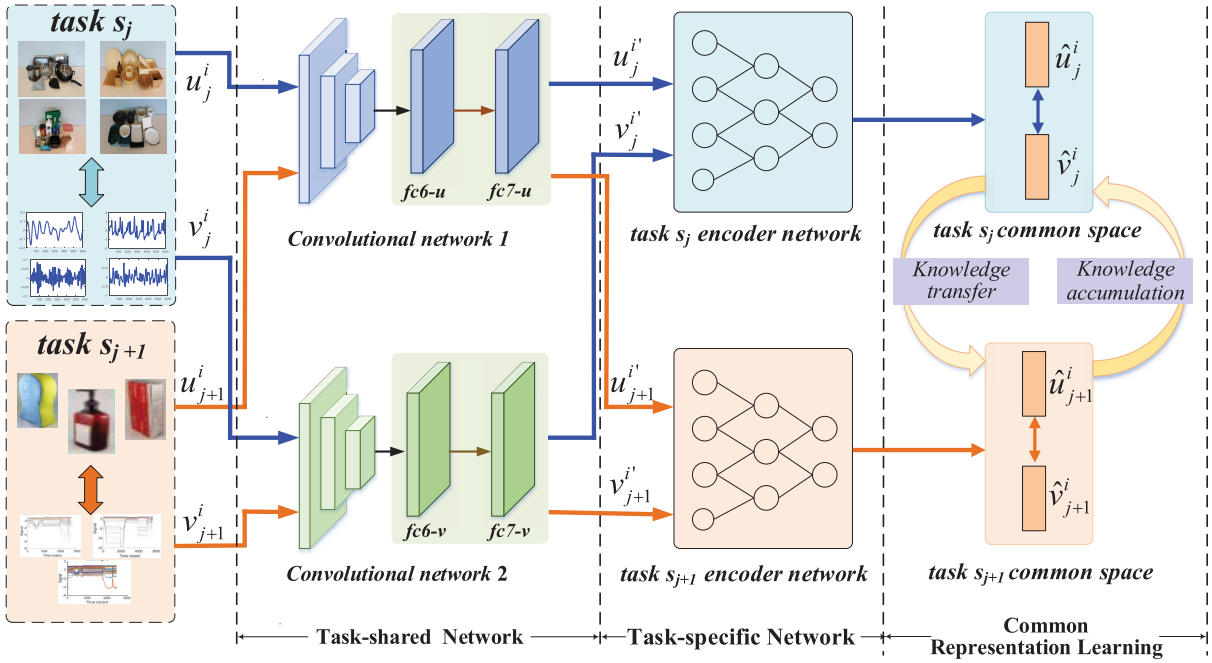


Fig. 2. Framework for lifelong cross-modal correlation learning.

What is more, the task of lifelong visual-tactile cross-modal learning faces two other challenges: knowledge correctness and knowledge application. In particular, knowledge correctness needs that the proposed model can achieve a good correlation of pairwise samples $\{(u_j^i, v_j^i), y_j^i\}$ in a single task S_j, which is a prerequisite for continuous correlation learning across different tasks. When the model performs well on the previous tasks, how to reuse the learned knowledge to help a new task learning is also critical for lifelong learning. As indicated earlier, the instance distribution of new task S_{j+1} is usually different from the previous task S_j. This makes it difficult to accumulate and transfer previously learned knowledge to increase its learning capacity. Moreover, it is likely to forget the knowledge learned from the previous task after the model is trained on a new task. It violates the principle of the definition of lifelong learning.

B. Challenges

From the earlier description, the lifelong visual-tactile cross-modal learning is a highly challenging task due to the following three major challenges.

1) *Modality Discrepancy:* Similar to other cross-modal learning, visual modality and tactile modality also exhibit high modality discrepancy, which are inconsistent on both distribution and representation. Therefore, it is difficult or even impossible to directly calculate their content similarity.
2) *Weakly Paired Matching:* Different from other cross-modal learning, visual-tactile cross-modal learning faces another challenge, which is weakly paired matching. As an existing method of cross-modal learning cannot fully utilize the weakly paired information, they may not be suitable for the task of visual-tactile cross-modal learning.
3) *Catastrophic Forgetting:* Lifelong learning is a key challenge for deep neural networks as they are vulnerable to the risk of catastrophic forgetting. It is mainly due to that the tasks in lifelong learning are usually from different instances of different domains. In particular, tactile signals usually have different representations in different tasks. This makes it difficult to transfer previously learned knowledge to new task learning. At the same time, the model trained on a new task usually forgets knowledge learned from previous tasks as catastrophic forgetting. However, it violates the nature of lifelong learning.

IV. Proposed Method

In this section, we first elaborate on the overall network architecture of the proposed LLCL model in Section IV-A and then detail the proposed model of LLCL IV-B. Finally, we give the optimization procedure of the model in Section IV-C.

A. Network Architecture

We construct a network architecture for lifelong cross-modal correlation learning, which is shown in Fig. 2. The network can be structurally divided into two subnetworks: 1) task-shared network, and 2) task-specific network. The parameters of task-shared network are shared for knowledge transfer and reuse in different tasks, whereas the task-specific network is used to ensure the extensibility and adaptability of the model to new tasks.

1) Task-Shared Network: As can be seen from Fig. 2, the task-shared network consists of two parallel network branches for visual modality and tactile modality, respectively. For the branch of visual modality, we adopt widely used VGG-16 as the basic model, except for the last fully connected

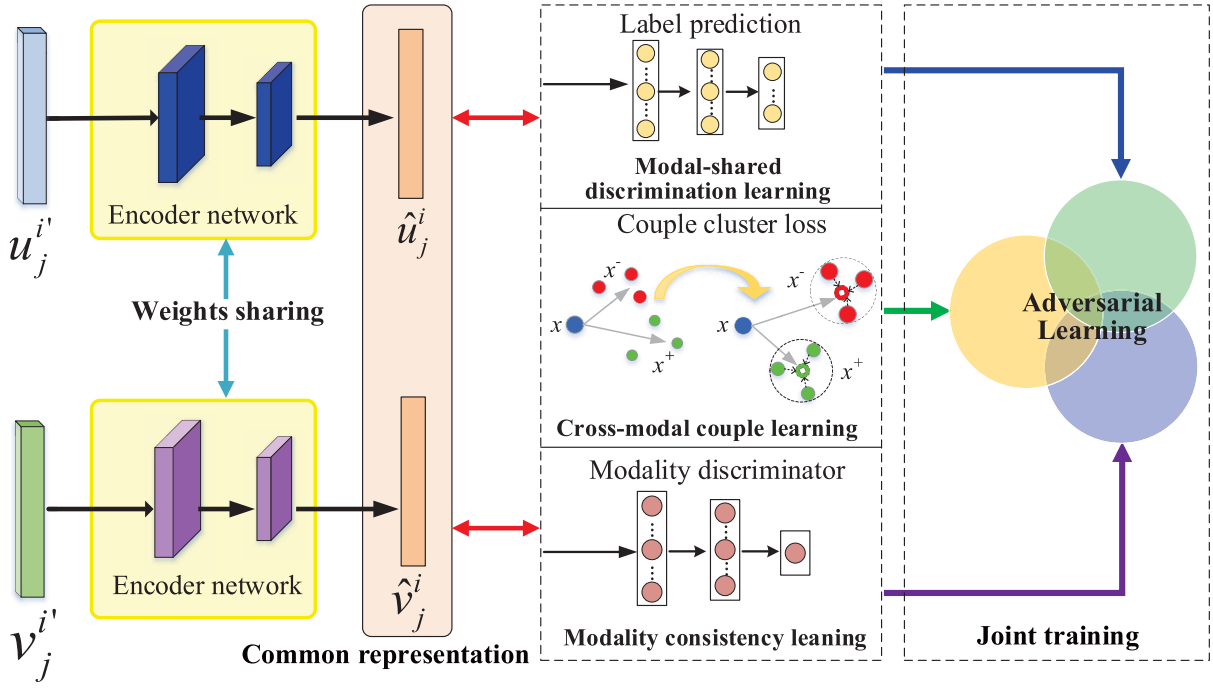


Fig. 3. Structure of intratask cross-modal correlation learning subnetwork.

layer. Here, the two 4096-dimensional full-connected layers are denoted as $fc6-u$ and $fc7-u$. Similar to the visual modality, the branch of tactile modality also uses the same network structure, namely convolutional layers and two full-connected layers $fc6-v$ and $fc7-v$. It is noted that the parameters in $fc7-u$ and $fc7-v$ of task-shared network are shared with each other, the purpose of which is to construct cross-modal correlation and share the high-level knowledge from different modalities.

2) Task-Specific Network: As shown in Fig. 3, the task-specific network also consists of two parallel branches, which are connected by some specified constraints on the common code layer. According to the modality category, the two branches of the task-specific network are concatenated to the two branches of the task-shared network. Finally, the representations learned from the task-shared network are fed into the task-specific network. The parameters of two parallel branches in the task-specific network are shared in the same task, while they are independent of each other between different tasks.

The proposed network architecture concatenating task-shared network and task-specific network can adaptively learn new tasks, which paves the way for lifelong visual-tactile cross-modal learning.

B. Lifelong Cross-Modal Learning

To realize LLCL, the LLCL model uses the task-shared network and task-specific network to, respectively, project d_u-dimensional image feature u_j^i and d_v-dimensional tactile feature v_j^i into a shared representation subspace

$$\hat{u}_j^i = f_u\left(u_j^i, \theta_s^u, \theta_j^u\right) : R^{d_u} \rightarrow R^d \quad (1)$$

$$\hat{v}_j^i = f_v\left(v_j^i, \theta_s^v, \theta_j^v\right) : R^{d_v} \rightarrow R^d \quad (2)$$

where $d \leq \min(d_u, d_v)$ is the dimension of common representation in the shared subspace, θ_s^u and θ_s^v denote the parameters of visual and tactile branches in the task-shared network, respectively, and θ_j^u and θ_j^v are the parameters of visual and tactile branches in the task-specific network for task S_j, respectively.

1) Intratask Cross-Modal Learning: First, we elaborate visual-tactile cross-modal learning of the proposed LLCL model in an individual task. As shown in Fig. 3, the common representations $\hat{u}_j^i$ and $\hat{u}_j^i$ of the two modalities are coupled at their code layer using semantic constraints. In this model, the semantic constraints consist of three optimization objectives: modal-shared discriminative learning, cross-modal coupled learning, and modal-adversarial consistency learning. The two concatenating networks with the semantic constraints form an end-to-end structure, which aims to learn discriminative common representations by joint training.

a) Modal-shared discriminative learning: Since the pairwise samples from visual and tactile modalities share the same semantic label, their consistency can be achieved by the modal-shared class label. To this end, a shared fully connected layer is used as the category classifier with softmax function. The optimization objective of modal-shared discriminative learning can be expressed as

$$L_{\text{dis}}\left(\theta_s, \theta_j, \theta_j^l\right) = \frac{1}{N}\sum\left(f_l\left(\hat{u}_j^i, y_j^i, \theta_j^l\right) + f_l\left(\hat{v}_j^i, y_j^i, \theta_j^l\right)\right) \quad (3)$$

where $\theta_s = (\theta_s^u, \theta_s^v)$, $\theta_j = (\theta_j^u, \theta_j^v)$, θ_j^l denotes the parameters of the shared classifier for task S_j, and $f_l(x, y, \theta)$ is the softmax loss defined as

$$f_l(x, y, \theta_l) = \sum_{k=1}^{c} 1\{y = k\} \log[\hat{p}(x, k, \theta_l)] \quad (4)$$

where $\hat{p}(x, k, \theta)$ is the probability distribution of categories of x over all classes.

By minimizing the term $L_{\text{dis}}(\theta_s, \theta_j, \theta_j^l)$, it makes the common representations have the optimal category discriminability. Namely, this optimization branch can ensure the learned representations from different classes staying apart, regardless of which modality the features come from.

b) Cross-modal coupled learning: The aforementioned optimization term mainly enables the common representations to have class discrimination within each modality. However, it is difficult to ensure the semantic correlation of paired samples. The cross-modal semantic correlation is the core of cross-modal learning. To this end, we proposed cross-modal couple learning to further improve the cross-modal semantic correlation with couple cluster constraint. It can simultaneously learn a feature center of each class for each modality and minimize the distance between the features and the pairwise class centers. Specifically, a sample from a specific class is used as positive sample of the class center denoted as x^+, while a sample from any other class is negative sample denoted as x^-. The positive sample x^+ class should be closer to the corresponding class centers c than any sample x^-, and the relative distance relationship should satisfy the following two constraints simultaneously:

$$\begin{aligned} L_{\text{cou}}^{u}(\theta_{sj}) &= \sum_{m,k} \max\left(0, l_2\left(\hat{v}_j^{m+} - c_u^m\right) + \delta - l_2\left(\hat{v}_j^{k*} - c_u^m\right)\right) \\ L_{\text{cou}}^{v}(\theta_{sj}) &= \sum_{m,k} \max\left(0, l_2\left(\hat{u}_j^{m+} - c_v^m\right) + \delta - l_2\left(\hat{u}_j^{k*} - c_v^m\right)\right) \end{aligned} \tag{5}$$

where $\theta_{sj} = (\theta_s, \theta_j)$, $l_2 = \|\cdot\|_2$ is L_2 norm, δ is the margin between positive and negative pairs, and $\hat{u}_j^{k*}$ and $\hat{v}_j^{k*}$ are, respectively, the nearest negative paired samples to the cluster centers c_v^m and c_u^m. The class center c_u^m and c_v^m can be calculated by

$$c_u^m = \frac{1}{N_u^m} \sum_{i}^{N_u^m} \hat{u}_j^{m+}, \quad c_v^m = \frac{1}{N_v^m} \sum_{i}^{N_v^m} \hat{u}_j^{m+} \tag{6}$$

where $\hat{u}_j^{m+}$ denotes the samples from the class m in visual modality, N_u^m is the number of corresponding samples, $\hat{v}_j^{m+}$ denotes the samples of the class m in tactile modality, and N_v^m is the number of corresponding samples.

Although (5) is similar to triplet constraint in the intuition form, our proposed cross-modal couple cluster constraint has the twofold advantages.

1) Different from randomly selecting a sample as an anchor of triple samples, cross-modal learning uses a cluster center to compute relative similarity relation of pairwise samples.
2) The optimization objective of cross-modal couple learning focuses on multiple samples rather than three samples in each iteration.

The above-mentioned improvement makes training more stable and accelerates convergence. Moreover, the selection of negative samples will further promote the correlation learning process.

The overall optimization objective of cross-modal couple learning can be modeled by integrating the above-mentioned two hinge loss, which is expressed as

$$L_{\text{cou}}(\theta_{sj}) = L_{\text{cou}}^{u}(\theta_{sj}) + L_{\text{cou}}^{v}(\theta_{sj}). \tag{7}$$

By minimizing $L_{\text{cou}}(\theta_{sj})$, it can ensure that cross-modal paired samples have compact group-invariant and semantic consistency, which is beneficial for weakly paired visual-tactile cross-modal learning.

c) Modal-adversarial consistency learning: In addition to intramodal discrimination learning and cross-modal couple learning, intermodal consistency is also essential for cross-modal learning. It is mainly due to that if visual and tactile modalities have consistent representations, the performance of cross-modal learning can be further enhanced. To this end, we propose modality consistency learning to reduce distribution discrepancy of representations between visual modality and tactile modality.

A modality discriminator is introduced for modal consistency learning. By confusing the modality source, the discriminator can reduces the distribution discrepancy of common representations. It consists of several full-connected layers with a softmax layer, whose parameters are denoted as θ_j^d. In the training process, all samples from tactile modality and visual modality are, respectively, assigned with the corresponding modality label. The optimization objective of modal consistency learning is expressed as

$$L_{\text{mod}}(\theta_{sj}, \theta_j^d) = -\frac{1}{N} \sum_{i=1}^{N} \left(f_{\text{mod}}(\hat{u}_j^i, m_i, \theta_j^d) + f_{\text{mod}}(\hat{v}_j^i, m_i, \theta_j^d)\right) \tag{8}$$

where m_i denotes the modality label of $\hat{u}_j^i$ or $\hat{v}_j^i$ and $f_{\text{mod}}(x, y, \theta)$ is the sigmoid cross-entropy loss, which is defined as

$$f_{\text{mod}}(x, y, \theta) = y \log \hat{p}(x, \theta) + (1 - y)\log(1 - \hat{p}(x, \theta)) \tag{9}$$

where $\hat{p}(x, \theta)$ is defined as

$$\hat{p}(x, \theta) = \frac{1}{1 + e^{-\phi(x,\theta)}}. \tag{10}$$

By maximizing $L_{\text{mod}}(\theta_{sj}, \theta_j^d)$, it can explicitly reduce the representations discrepancy between these two modalities and further improve the performance of common representation learning.

2) Intertask Knowledge Distillation: Correctness of knowledge is the basis and premise of lifelong learning. The intratask cross-modal correlation learning aims to make the proposed model achieve the optimal performance on each single cross-modal learning task, which ensures that the knowledge learned gained from older tasks is correct. Our final goal is to incrementally learn a sequence of cross-modal learning tasks and use the knowledge learned from the past tasks to help perform new tasks and also learn from it. As mentioned earlier, when the model learned from past tasks adapts a new task, it usually suffers from catastrophic forgetting.

Recently, Li and Hoiem [59] proposed to use knowledge distillation [63] to tackle catastrophic forgetting and achieved

excellent performance in some classification tasks. Moreover, the method does not need training data from the previous tasks during learning a new task. Inspired by the method, we introduce the idea of knowledge distillation into lifelong visual-tactile cross-modal learning.

Specifically, we first record the output probability distribution $\hat{y}_j$ of samples from the new task S_{j+1} using the learned parameters (θ_s, θ_j) in the task S_j network. Then, we use knowledge distillation loss to enforce the probability distribution y_j of the new task S_{j+1} in the new parameters θ_s and the original parameters θ_j close to $\hat{y}_j$ during learning parameters (θ_s, θ_{j+1}) for the new task S_{j+1}. The optimization objective for intertask knowledge distillation is expressed as

$$L_{\text{ter}}(\theta_s, \theta_{j+1}) = -\sum_{b=1}^{N} y_j \log \hat{y}_j. \tag{11}$$

By minimizing $L_{\text{ter}}(\theta_s, \theta_{j+1})$, the performance of the model on the old task S_j can be maintained during knowledge learned from the task S_j is transferred into the task S_{j+1} and help it learn better. Finally, the learned model can work well on both previous task S_j and new task S_{j+1}.

In a nutshell, with the intertask knowledge distillation, we can reuse the knowledge learned from the past task, to make the model achieve better performance on new cross-modal learning tasks, while preserving or even further enhancing the correlation learning capabilities of the model on previous tasks. Eventually, we can perform lifelong visual-tactile cross-modal learning.

C. Algorithm Optimization

From the earlier discussions, the optimization process of the proposed model can be divided into two stages: 1) network optimization of task $S_0 (j = 0)$ focuses on cross-modal correlation learning of visual modality and tactile modality and 2) network optimization of task $S_j (j \geq 1)$ emphasizes the knowledge transfer and accumulation between a sequence of visual-tactile cross-modal learning tasks.

1) Stage 1: Network Optimization of Task $S_0 (j = 0)$: For the initial task S_0, the network parameters θ_s and θ_0 are jointly trained by the (3), (7), and (8). The formally optimization objective of the model can be expressed as

$$L_0(\theta_{s0}, \theta_0^l, \theta_0^d) = L_{\text{dis}}(\theta_{s0}, \theta_0^l) + L_{\text{cou}}(\theta_{s0}) - L_{\text{mod}}(\theta_{s0}, \theta_0^d). \tag{12}$$

We seek the optimal parameters θ_{s0}, θ_0^l and θ_0^d by minimizing the loss functions $L_{\text{dis}}(\theta_{s0}, \theta_0^l)$, $L_{\text{cou}}(\theta_{s0})$, and $L_{\text{ter}}(\theta_{s0})$ while simultaneously maximizing the loss function $L_{\text{mod}}(\theta_{s0})$ in the training process, which is an adversarial training strategy [44].

The optimization process of (12) can be described with two concurrent subprocesses as follows:

$$(\hat{\theta}_{s0}, \hat{\theta}_0^l) = \arg\min_{\theta_{s0}, \theta_l} L_0(\theta_{s0}, \theta_0^l, \hat{\theta}_0^d) \tag{13}$$

$$(\hat{\theta}_0^d) = \arg\max_{\theta_0^d} L_0(\hat{\theta}_{s0}, \hat{\theta}_0^l, \theta_0^d). \tag{14}$$

Clearly, (13) and (14) have opposite optimization objectives as minimax game of two subprocesses. As proposed in [64], the optimization of (12) can be implemented through the adversarial learning strategy. By adding a gradient reversal layer (GRL) between the encoder network and the modality discriminator, the objective functions of (13) and (14) can be optimized simultaneously with stochastic gradient descent (SGD) optimization algorithms.

Algorithm 1 Optimization Process of the LLCL Model for Task $S_j (j \geq 1)$

Require:
Task-shared parameters θ_s learned from task S_0 to task S_j, task-specific parameters θ_j and θ_j^l for task S_j; cross-modal visual and tactile training data and their category labels for task S_{j+1}, which is denoted as a_{j+1}^i.

Ensure:
New task-share parameters θ_s learned from task S_{j+1}, and task-specific parameters θ_{j+1} and θ_{j+1}^l for task S_{j+1}.

1: Calculate the probability distribution of training data from task S_{j+1} on the model of task S_j with parameters θ_s, θ_j and θ_j^l: $Y_o \leftarrow Net(a_{j+1}^i, \theta_s, \theta_j, \theta_j^l)$.
2: Initialize the task-specific parameters θ_{j+1} and θ_{j+1}^l for task S_{j+1}: $\theta_{j+1}, \theta_{j+1}^l \leftarrow Randinit(|\theta_{j+1}|, |\theta_{j+1}^l|)$.
3: **for** number of iterations **do**
4: Update parameters θ_{sj+1}, θ_{j+1}^l and θ_{j+1}^d with Eq.(15) by adversarial learning:
5: $(\theta_{sj+1}, \theta_{j+1}^l, \theta_{j+1}^d) \leftarrow argmaxmin(L_{j+1}(\theta_{sj+1}, \theta_{j+1}^l, \theta_{j+1}^d))$;
6: **end for**
7: **return** $\theta_{sj+1}, \theta_{j+1}^l, \theta_{j+1}^d$.

2) Stage 2: Network Optimization of Task $S_j (j \geq 1)$: When new tasks S_{j+1} arriving in sequence, the lifelong visual-tactile cross-modal learning is realized by joint optimization of intratask cross-modal correlation learning and intertask knowledge distillation. The network parameters θ_s and θ_j are jointly trained by (3), (7), (8), and (10). The formally optimization objective of the model can be expressed as

$$\begin{aligned} L_{j+1}(\theta_{sj+1}, \theta_{j+1}^l, \theta_{j+1}^d) = {} & L_{\text{dis}}(\theta_{sj+1}, \theta_{j+1}^l) + L_{\text{cou}}(\theta_{sj+1}) \\ & - L_{\text{mod}}(\theta_{sj+1}, \theta_{j+1}^d) + L_{\text{ter}}(\theta_{sj+1}, \theta_j^l) \end{aligned} \tag{15}$$

where $\theta_{sj+1} = (\theta_s, \theta_{j+1})$.

Similar to the network optimization process of task S_0, the optimization objective functions of network optimization of task $S_j (j \geq 1)$ can be expressed as

$$(\hat{\theta}_{sj}, \hat{\theta}_j^l) = \arg\min_{\theta_v, \theta_t, \theta_l} L(\theta_{sj+1}, \theta_{j+1}^l, \hat{\theta}_{j+1}^d) \tag{16}$$

$$(\hat{\theta}_j^d) = \arg\max_{\theta_j^d} L(\hat{\theta}_{sj+1}, \hat{\theta}_{j+1}^l, \theta_{j+1}^d). \tag{17}$$

Overall, the optimization process of the proposed LLCL model is summarized in Algorithm 1 when task $S_j (j \geq 1)$ arrives.

V. Experiments

The proposed lifelong cross-modal correlation learning is evaluated on three publicly available data sets. First, the adopted data sets, evaluation metric, comparison methods,

and implementation details are introduced. Then, experiment results and their analysis are given. Finally, we further analyze the effectiveness of lifelong learning strategy in the model.

A. Data Set

We conduct experiments of lifelong cross-modal correlation learning on three publicly available data sets, namely LTM_108 [67], LTM_96 [68] and PFN_VT [26]. The detail and splits of the data sets are introduced as follows.

1) *LTM_108:* It has 108 different surface material instances, which are divided into nine categories based on the physical attribute. The data set provides the visual images and the tactile acceleration traces for each surface instance. Different from the instance recognition of the work [34], we use this data set to tackle the more challenge cross-modal learning. Therefore, we need to reorganize the data set. Following the reorganization way [11], there are 90×19 sample pairs of visual and tactile for training, 108 tactile sample as query set, and 90 image for retrieval set.
2) *LTM_96:* It includes 96 different surface material instances from nine categories of physical attribute. This data set adopts a similar setup with LTM_108 to collect visual images and tactile data, but their process of data collection is different. They have different surface instances except a few overlapping. In this data set, we adopt the same reorganization way with LTM_108. Eventually, there are training set with 78×19 visual/tactile pairs, query set with 96 tactile samples, and retrieval set with 78 images.
3) *PFN_VT:* It is released by preferred networks. It consist of 25 surface materials with different textures, where 15 materials are used as training, while the remaining 10 are used as unknown materials to evaluate the cross-modal learning model. In this original data set, each instance has ten images and ten tactile sequences of each surface material. We randomly select eight image-tactile sequence pairs as training data, and the remaining two image-tactile sequence pairs are used as testing data of known materials. Then, we randomly select two pairs from 20 image-tactile sequence pairs of ten unknown materials, which are used as the testing data of unknown materials. Thus, there are 120 image-tactile pairs for training, and 50 image-tactile pairs for testing, where the 50 images as retrieval set and 50 tactile sequences as a query set. Similar to [26], we augment our training data and obtain 960 training sample pairs.

In the three data sets, the tactile signals are three-axis time/force sequences. In this article, these tactile sequences of three dimensions are converted into one dimension using the DFT321 algorithm [69]. Then, we use the short-time Fourier transform (STFT) to convert the tactile signal into spectrogram images as the input data of tactile modality for the proposed model.

Please note that the sequence of the tasks to be learned has a significant effect on the performance of the model. To consider the sequence effect, we perform extensive experiments for all possible sequences. The experimental results show that the proposed modal can achieve the best performance for all task in the sequence as LTM_108 $\rightarrow$ LTM_96 $\rightarrow$ PFN_VT, which is consistent with the idea of knowledge distillation. The main reason is that knowledge learned from a large data set has better generalization, which can help small data set to learn an effective model. Therefore, in this LLCL setting, we assume that the training data of three different tasks arrive in sequence as LTM_108 $\rightarrow$ LTM_96 $\rightarrow$ PFN_VT, and training samples in previous tasks are not available when learning new tasks.

B. Compared Methods and Evaluation Metric

To verify the effectiveness of our proposed LLCL method, we compare it with eight methods on the LMT data set. The comparison methods are CCA [33], MCCA [65], CCCA [65], WMCA [24], GMMFA [66], DCCA [38], DAML [45], and ACMR [46].

Similar to [13], mean average precision (MAP) is used as the evaluation metric to compare the retrieval performance in this experiment. Besides MAP, we also use the Top_k accuracy for performance evaluation, which denotes the average accuracy of the first k returned results of all query samples. For more comprehensive evaluation, precision–scope (PS) curves and precision–recall (PR) curves of different methods will be given.

C. Implementation Details

In the proposed LLCL model, we use a two-layer fully connected network as a task-specific network. In particular, the network structure is 2048 $\rightarrow$ 1024, whose parameters are shared. A three fully connected layers are used as the modality discriminator, where the number of nodes in the first two layers is 1024 $\rightarrow$ 512 $\rightarrow$ 2, and the node number of the last layer is the category number of the specific task.

In training time, the minibatch size is set to 64, the learning late μ is 0.01, and Adaptive Moment Estimation (Adam) [70] is selected as optimization algorithms. During training of the model, we note that a strong modal-adversarial learning degenerates the performance of the model on the contrary. To mitigate the issue, the parameters of modality discriminator are updated once when updating the parameters of representation learning five times in these experiments.

D. Experimental Results and Analysis

In this section, we present the experimental results, which contains the results of our proposed LLCL model under a lifelong setting and the results trained on the corresponding task individually, denoted as LLCL_s. It aims to verify the effectiveness of the proposed model in a single task by comparing it with other methods and highlight the role of lifelong learning in the proposed model. The results on three visual-tactile cross-modal tasks are summarized in Table I.

From Table I, it can be seen that our proposed model LLCL and LLCL_s achieves the higher MAP, Top_1, and Top_5 compared with other methods, and there are the same trends on the three tasks. It shows that the proposed method is effective in visual-tactile cross-modal learning.

TABLE I
RETRIEVAL RESULTS OF DIFFERENT METHODS

Model	LTM_108			LTM_69			PFN_VT		
	MAP	Top_1	Top_5	MAP	Top_1	Top_5	MAP	Top_1	Top_5
CCA [33]	23.2%	33.3%	25.5%	24.8%	35.1%	27.9%	24.9%	27.3%	18.2%
MCCA [68]	16.9%	8.01%	14.4%	18.6%	12.9%	16.3%	22.7%	24.3%	16.2%
CCCA [68]	16.4%	13.9%	14.3%	17.9%	15.6%	17.2%	29.5%	33.5%	22.8%
WMCA [24]	16.2%	17.6%	14.4%	17.7%	21.5%	19.8%	28.2%	28.5%	19.8%
GMMFA [69]	42.8%	38.9%	35.4%	43.1%	39.5%	37.8%	36.2%	39.5%	30.2%
DCCA [38]	31.2%	35.0%	34.6%	33.9%	38.1%	36.2%	41.6%	42.1%	25.9%
DAML [45]	56.3%	60.7%	58.2%	57.2%	61.8%	58.1%	55.2%	55.9%	43.1%
ACMR [46]	59.5%	63.0%	61.1%	61.6%	65.8%	63.2%	57.9%	58.8%	45.2%
LLCL_s	62.6%	66.8%	63.7%	63.3%	68.6%	65.4%	59.6%	62.2%	58.1%
LLCL	64.8%	69.5%	66.7%	65.2%	70.8%	67.6%	62.9%	64.5%	60.3%

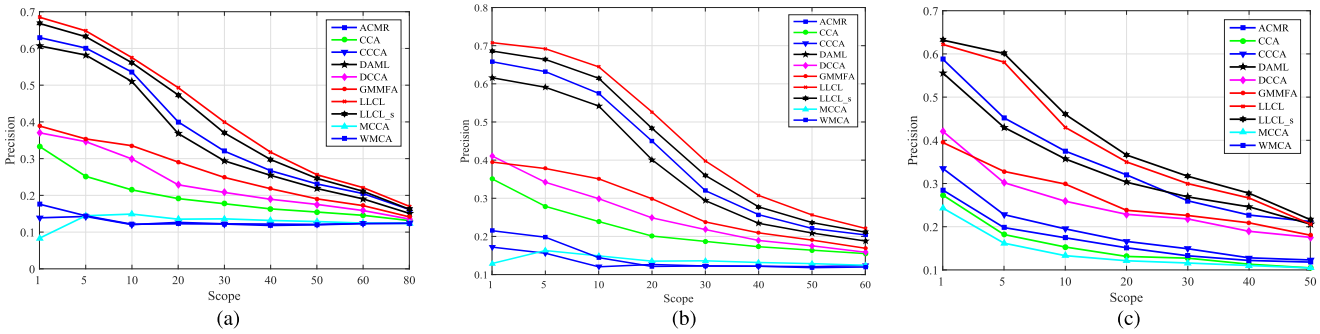


Fig. 4. PS curves of different methods on different data sets. (a) PS curves on LTM_108. (b) PS curves on LTM_96. (c) PS curves on PFN_VT.

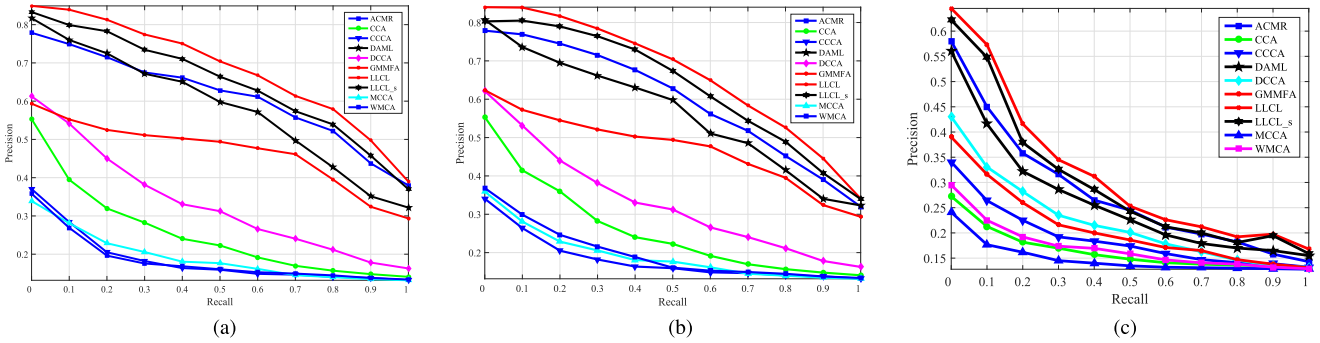


Fig. 5. PR curves of different methods on different data sets. (a) PR curves on LTM_108. (b) PR curves on LTM_96. (c) PR curves on PFN_VT.

Furthermore, we can observe that DAML and ACMR achieve the better performance among all the compared methods. It is mainly because they incorporate intramodality discrimination learning and intermodality correlation learning into a unified deep network to learn common representation with adversarial training. Comparing LLCL-s with DAML and ACMR, the performance of LLCL-s is significantly superior to them. The essential difference between these methods is that DAML and ACMR, respectively, use contrastive constraints and triplet constraints learning cross-modal correlation, while our proposed LLCL_s imposes couple cluster constraints to preserve the pairwise correlation in adversarial representation leaning. It highlights the importance of propose cross-modal couple cluster constraints in the weakly paired visual-tactile cross-modal learning.

In particular, it is clear from Table I that LLCL outperforms LLCL_s. LLCL_s is to learn a specific model for each individual task, while LLCL aims to l a model for all tasks by transferring knowledge among different tasks. The results demonstrate that with the lifelong learning strategy, the proposed LLCL model can improve the performance of the model in a new task and further enhance the correlation learning capabilities of the model on previous tasks.

For more comprehensive analysis, the PS curves and PR curves of all methods on the three data sets are shown in Figs. 4 and 5, respectively. From the PS curves of all methods in Fig. 4, it can be seen that the proposed LLCL_s model and LLCL model can search more correct visual samples in the same returned results for all tasks. The results in Fig. 5 show that the PR curves of the proposed LLCL_s model and LLCL model are almost always above the compared methods. It indicates that the performances of all compared methods are inferior to the proposed method for the task of visual-tactile cross-modal learning.

TABLE II
COMPARISONS WITH FINE-TUNING

Model	LTM_108			LTM_69			PFN_VT		
	MAP	Top_1	Top_5	MAP	Top_1	TOP_5	MAP	Top_1	Top_5
LLCL(LTM_108)	62.6 %	69.5%	63.7%	—	—	—	—	—	—
LLCL($LTM\_108 \rightarrow LTM\_96$)	65.2 %	71.2%	68.3%	65.8%	71.2%	68.4%	—	—	—
Ft($LTM\_108 \rightarrow LTM\_96$)	35.1%	40.2%	32.6%	64.3%	68.6%	64.5%	—	—	—
LLCL($LTM\_108 + LTM\_96 \rightarrow PFN\_VT$)	64.8%	69.5%	66.8%	65.2%	70.8%	67.6%	62.9%	64.5%	60.3%
Ft($LTM\_108 + LTM\_96 \rightarrow PFN\_VT$)	17.7%	24.8%	16.1%	37.2%	41.1%	36.4%	61.2%	62.8%	58.6%

E. Effectiveness of Lifelong Learning Strategy

To study the effectiveness of knowledge distillation in the proposed LLCL model, we compare the proposed lifelong learning strategy with a traditional fine-tuning method. For a fair comparison, we directly remove the term the intertask knowledge distillation from 15 and remain other terms for fine-tuning. Since we assume that the three tasks arrive in sequence as LTM_108 $\rightarrow$ LTM_96 $\rightarrow$ PFN_VT, we first use the data from LTM_96 to fine-tune the model learned from LTM_108, which is denoted as Ft(LTM_108 $\rightarrow$ LTM_96), where the data of PFN_VT is not available under this process. Then, we further fine-tune for PFN_VT on the model trained from LTM_108 and LTM_96, denoted as Ft(LTM_108 + LTM_96 $\rightarrow$ PFN_VT). The detailed experimental results are shown in Table II.

From the results in Table II, it can be observed that the performance of the model on the past tasks would significantly degrade after fine-tuning on the data from a new task, while our proposed model can preserve or even improve the performance of the model on old tasks. It demonstrates that catastrophic forgetting is a major problem for tasks of lifelong learning, the proposed lifelong learning strategies based on intertask knowledge distillation can maintain the ability of the learned model to perform past learned tasks when it is adapted new tasks. Moreover, our proposed LLCL model is also superior to the performance of a fine-tuned model on new tasks, which verifies that the proposed lifelong learning method can effectively transfer knowledge among tasks to enhance the performance of model on new tasks.

Besides, it is seen that its performance on the LTM_108 increases when the proposed model is transferred from LTM_108 to LTM_96, while its performance on LTM_108 degrades when it is transferred from LTM_108 + LTM_96 to PFN_VT. It is mainly due to that the task LTM_108 and the task LTM_96 are highly correlated and have similar distributions, and continual learning will improve the performance of the model, while there is large distribution discrepancy between the task LTM_108 and the task LTM_96, continual learning on the contrary may degrade its performance. It is consistent with the nature of catastrophic forgetting of deep neural network.

VI. CONCLUSION

In this article, the framework of visual-tactile cross-modal is proposed for robotic material perception. Crucially, we proposed to investigate visual-tactile cross-modal learning in the lifelong learning framework. Moreover, we develop a lifelong cross-modal correlation learning model. Experimental results on the three publicly available data sets demonstrate the validity of the proposed method. The framework and model provide a new viewpoint for remote material perception, particularly in remote unstructured environments. Since the task of lifelong visual-tactile cross-modal material perception is very complicated, it is associated with many different challenges. The focus of this article is to study the algorithm of lifelong visual-tactile cross-modal learning, and other aspects still need to be further explored. For example, we did not consider seamlessly adding new tasks into the learned model in this article. In addition, the model cannot adaptively expand the network when a new task arrives. Therefore, the method currently cannot be used in practical robotic applications. In the future, the spiking neural networks may be used for practical implementation [71]. In addition, we will aim to resolve the above-mentioned limitations and construct real robot platform for experimental verification.

REFERENCES

[1] Z. Li, B. Huang, Z. Ye, M. Deng, and C. Yang, "Physical human–robot interaction of a robotic exoskeleton by admittance control," *IEEE Trans. Ind. Electron.*, vol. 65, no. 12, pp. 9614–9624, Dec. 2018.

[2] W. Yuan, Y. Mo, S. Wang, and E. H. Adelson, "Active clothing material perception using tactile sensing and deep learning," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 1–8.

[3] H. Liu, Y. Wu, and F. Sun, "Extreme trust region policy optimization for active object recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 6, pp. 2253–2258, Jun. 2018.

[4] Q. Dang, J. Yin, B. Wang, and W. Zheng, "Deep learning based 2D human pose estimation: A survey," *Tsinghua Sci. Technol.*, vol. 24, no. 6, pp. 663–676, Dec. 2019.

[5] Z. Li, B. Huang, A. Ajoudani, C. Yang, C.-Y. Su, and A. Bicchi, "Asymmetric bimanual control of dual-arm exoskeletons for human-cooperative manipulations," *IEEE Trans. Robot.*, vol. 34, no. 1, pp. 264–271, Feb. 2018.

[6] H. Liu, F. Sun, and X. Zhang, "Robotic material perception using active multimodal fusion," *IEEE Trans. Ind. Electron.*, vol. 66, no. 12, pp. 9878–9886, Dec. 2019.

[7] Z. Li, C. Deng, and K. Zhao, "Human-cooperative control of a wearable walking exoskeleton for enhancing climbing stair activities," *IEEE Trans. Ind. Electron.*, vol. 67, no. 4, pp. 3086–3095, Apr. 2020.

[8] Z. Li, J. Li, S. Zhao, Y. Yuan, Y. Kang, and C. P. Chen, "Adaptive neural control of a kinematically redundant exoskeleton robot using brain–machine interfaces," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 12, pp. 3558–3571, Dec. 2019.

[9] Z. Li, C. Xu, Q. Wei, C. Shi, and C.-Y. Su, "Human-inspired control of dual-arm exoskeleton robots with force and impedance adaptation," *IEEE Trans. Syst., Man, Cybern. Syst.*, pp. 1–10, 2019.

[10] H. Liu, Y. Yu, F. Sun, and J. Gu, "Visual–tactile fusion for object recognition," *IEEE Trans. Autom. Sci. Eng.*, vol. 14, no. 2, pp. 996–1008, Apr. 2017.

[11] W. Zheng, H. Liu, B. Wang, and F. Sun, "Cross-modal surface material retrieval using discriminant adversarial learning," *IEEE Trans. Ind. Informat.*, vol. 15, no. 9, pp. 4978–4987, Sep. 2019.

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

[13] W. Zheng, H. Liu, B. Wang, and F. Sun, "Cross-modal material perception for novel objects: A deep adversarial learning method," *IEEE Trans. Automat. Sci. Eng.*, Oct. 9, 2019, early access, doi: 10.1109/TASE.2019.2941230.

[14] E. Kerr, T. McGinnity, and S. Coleman, "Material recognition using tactile sensing," *Expert Syst. Appl.*, vol. 94, pp. 94–111, Mar. 2018.

[15] E. Baumgartner, C. B. Wiebel, and K. R. Gegenfurtner, "Visual and haptic representations of material properties," *Multisensory Res.*, vol. 26, no. 5, pp. 429–455, 2013.

[16] R. W. Fleming, "Visual perception of materials and their properties," *Vis. Res.*, vol. 94, pp. 62–75, Jan. 2014.

[17] H. Qiao, M. Wang, J. Su, S. Jia, and R. Li, "The concept of 'attractive region in environment' and its application in high-precision tasks with low-precision systems," *IEEE/ASME Trans. Mechatronics*, vol. 20, no. 5, pp. 2311–2327, Oct. 2015.

[18] R. Camoriano, G. Pasquale, C. Ciliberto, L. Natale, L. Rosasco, and G. Metta, "Incremental robot learning of new objects with fixed update time," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2017, pp. 3207–3214.

[19] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter, "Continual lifelong learning with neural networks: A review," *Neural Netw.*, vol. 113, pp. 54–71, May 2019.

[20] M. Dehghan, Z. Zhang, M. Siam, J. Jin, L. Petrich, and M. Jagersand, "Online object and task learning via human robot interaction," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2019, pp. 2132–2138.

[21] S. Thrun and T. M. Mitchell, "Lifelong robot learning," *Robot. Auton. Syst.*, vol. 15, nos. 1–2, pp. 25–46, Jul. 1995.

[22] T. Lesort, V. Lomonaco, A. Stoian, D. Maltoni, D. Filliat, and N. Díaz-Rodríguez, "Continual learning for robotics: Definition, framework, learning strategies, opportunities and challenges," 2019, *arXiv:1907.00182*. [Online]. Available: http://arxiv.org/abs/1907.00182

[23] H. Liu, F. Sun, and B. Fang, "Lifelong learning for heterogeneous multimodal tasks," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2019, pp. 6158–6164.

[24] O. Kroemer, C. H. Lampert, and J. Peters, "Learning dynamic tactile sensing with robust vision-based training," *IEEE Trans. Robot.*, vol. 27, no. 3, pp. 545–557, Jun. 2011.

[25] P. Falco, S. Lu, A. Cirillo, C. Natale, S. Pirozzi, and D. Lee, "Cross-modal visuo-tactile object recognition using robotic active exploration," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2017, pp. 5273–5280.

[26] K. Takahashi and J. Tan, "Deep visuo-tactile learning: Estimation of tactile properties from images," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2019, pp. 8951–8957.

[27] S. Luo, W. Yuan, E. Adelson, A. G. Cohn, and R. Fuentes, "ViTac: Feature sharing between vision and tactile sensing for cloth texture recognition," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 2722–2727.

[28] J. Lin, R. Calandra, and S. Levine, "Learning to identify object instances by touch: Tactile recognition via multimodal matching," 2019, *arXiv:1903.03591*. [Online]. Available: http://arxiv.org/abs/1903.03591

[29] W. Yuan, S. Wang, S. Dong, and E. Adelson, "Connecting look and feel: Associating the visual and tactile properties of physical materials," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5580–5588.

[30] H. Liu, F. Wang, F. Sun, and B. Fang, "Surface material retrieval using weakly paired cross-modal learning," *IEEE Trans. Autom. Sci. Eng.*, vol. 16, no. 2, pp. 781–791, Apr. 2019.

[31] Y. Li, J.-Y. Zhu, R. Tedrake, and A. Torralba, "Connecting touch and vision via cross-modal prediction," 2019, *arXiv:1906.06322*. [Online]. Available: http://arxiv.org/abs/1906.06322

[32] Y. Yu, S. Tang, K. Aizawa, and A. Aizawa, "Category-based deep CCA for fine-grained venue discovery from multimodal data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 4, pp. 1250–1258, Apr. 2019.

[33] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Comput.*, vol. 16, no. 12, pp. 2639–2664, Dec. 2004.

[34] N. Rasiwasia *et al.*, "A new approach to cross-modal multimedia retrieval," in *Proc. Int. Conf. Multimedia (MM)*, 2010, pp. 251–260.

[35] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik, "A multi-view embedding space for modeling Internet images, tags, and their semantics," *Int. J. Comput. Vis.*, vol. 106, no. 2, pp. 210–233, Jan. 2014.

[36] D. Li, N. Dimitrova, M. Li, and I. K. Sethi, "Multimedia content processing through cross-modal association," in *Proc. 11th ACM Int. Conf. Multimedia (MULTIMEDIA)*, 2003, pp. 604–611.

[37] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2011, pp. 689–696.

[38] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1247–1255.

[39] W. Wang, R. Arora, K. Livescu, and J. Bilmes, "On deep multi-view representation learning," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1083–1092.

[40] F. Feng, R. Li, and X. Wang, *Deep Correspondence Restricted Boltzmann Machine for Cross-Modal Retrieval*. Amsterdam, The Netherlands: Elsevier, 2015.

[41] J. Tang, J. Lin, Z. Li, and J. Yang, "Discriminative deep quantization hashing for face image retrieval," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 12, pp. 6154–6162, Dec. 2018.

[42] X. Zhe, S. Chen, and H. Yan, "Deep class-wise hashing: Semantics-preserving hashing via class-wise loss," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, doi: 10.1109/TNNLS.2019.2921805.

[43] X. Yang, P. Zhou, and M. Wang, "Person reidentification via structural deep metric learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 10, pp. 2987–2998, Oct. 2019.

[44] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[45] X. Xu, L. He, H. Lu, L. Gao, and Y. Ji, "Deep adversarial metric learning for cross-modal retrieval," *World Wide Web*, vol. 22, no. 2, pp. 657–672, Mar. 2019.

[46] B. Wang, Y. Yang, X. Xu, A. Hanjalic, and H. T. Shen, "Adversarial cross-modal retrieval," in *Proc. ACM Multimedia Conf. (MM)*, 2017, pp. 154–162.

[47] H. He, Z. Ni, and J. Fu, "A three-network architecture for on-line learning and optimization based on adaptive dynamic programming," *Neurocomputing*, vol. 78, no. 1, pp. 3–13, Feb. 2012.

[48] G. I. Parisi, J. Tani, C. Weber, and S. Wermter, "Lifelong learning of human actions with deep neural network self-organization," *Neural Netw.*, vol. 96, pp. 137–149, Dec. 2017.

[49] X. Su, S. Guo, T. Tan, and F. Chen, "Generative memory for lifelong learning," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, doi: 10.1109/TNNLS.2019.2927369.

[50] H. He, S. Chen, K. Li, and X. Xu, "Incremental learning from stream data," *IEEE Trans. Neural Netw.*, vol. 22, no. 12, pp. 1901–1914, Dec. 2011.

[51] S. Hou, X. Pan, C. Change Loy, Z. Wang, and D. Lin, "Lifelong learning via progressive distillation and retrospection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 437–452.

[52] L. Shao, F. Zhu, and X. Li, "Transfer learning for visual categorization: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 5, pp. 1019–1034, May 2015.

[53] H. Qiao, X. Xi, Y. Li, W. Wu, and F. Li, "Biologically inspired visual model with preliminary cognition and active attention adjustment," *IEEE Trans. Cybern.*, vol. 45, no. 11, pp. 2612–2624, Nov. 2015.

[54] H. Qiao, Y. Li, F. Li, X. Xi, and W. Wu, "Biologically inspired model for visual cognition achieving unsupervised episodic and semantic feature learning," *IEEE Trans. Cybern.*, vol. 46, no. 10, pp. 2335–2347, Oct. 2016.

[55] Z. Ding and Y. Fu, "Deep transfer low-rank coding for cross-domain learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 6, pp. 1768–1779, Jun. 2019.

[56] J. Ramapuram, M. Gregorova, and A. Kalousis, "Lifelong generative modeling," 2017, *arXiv:1705.09847*. [Online]. Available: http://arxiv.org/abs/1705.09847

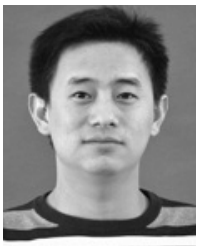
[57] H. Shin, J. K. Lee, J. Kim, and J. Kim, "Continual learning with deep generative replay," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 2990–2999.

[58] A. A. Rusu *et al.*, "Progressive neural networks," 2016, *arXiv:1606.04671*. [Online]. Available: http://arxiv.org/abs/1606.04671

[59] Z. Li and D. Hoiem, "Learning without forgetting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 12, pp. 2935–2947, Dec. 2018.

[60] K. James *et al.*, "Overcoming catastrophic forgetting in neural networks," *Proc. Nat. Acad. Sci. USA*, vol. 114, no. 13, pp. 3521–3526, Mar. 2017.

[61] S.-A. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert, "ICaRL: Incremental classifier and representation learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2001–2010.
[62] A. Rannen, R. Aljundi, M. B. Blaschko, and T. Tuytelaars, "Encoder based lifelong learning," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1320–1328.
[63] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*. [Online]. Available: http://arxiv.org/abs/1503.02531
[64] Y. Ganin *et al.*, "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2030–2096, May 2015.
[65] N. Rasiwasia, D. Mahajan, V. Mahadevan, and G. Aggarwal, "Cluster canonical correlation analysis," in *Proc. Artif. Intell. Statist.*, 2014, pp. 823–831.
[66] A. Sharma, A. Kumar, H. Daume, and D. W. Jacobs, "Generalized multiview analysis: A discriminative latent space," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2160–2167.
[67] M. Strese, Y. Boeck, and E. Steinbach, "Content-based surface material retrieval," in *Proc. IEEE World Haptics Conf. (WHC)*, Jun. 2017, pp. 352–357.
[68] M. Strese, C. Schuwerk, A. Iepure, and E. Steinbach, "Multimodal feature-based surface material classification," *IEEE Trans. Haptics*, vol. 10, no. 2, pp. 226–239, Apr. 2017.
[69] N. Landin, J. M. Romano, W. McMahan, and K. J. Kuchenbecker, "Dimensional reduction of high-frequency accelerations for haptic rendering," in *Proc. Int. Conf. Hum. Haptic Sens. Touch Enabled Comput. Appl.* Springer, 2010, pp. 79–86.
[70] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," Sep. 2015, *arXiv:1412.6980*. [Online]. Available: https://arxiv.org/abs/1412.6980
[71] J. Han, Z. Li, W. Zheng, and Y. Zhang, "Hardware implementation of spiking neural networks on FPGA," *Tsinghua Sci. Technol.*, vol. 25, no. 4, pp. 479–486, Aug. 2020.



Wendong Zheng is currently pursuing the Ph.D. degree in electrical engineering with the State Key Laboratory of Reliability and Intelligence of Electrical Equipment, School of Electrical Engineering, Hebei University of Technology, Tianjin, China.

His research interests include cross-modal retrieval, robot perception, and learning.

Huaping Liu (Senior Member, IEEE) is currently an Associate Professor with the Department of Computer Science and Technology, Tsinghua University, Beijing, China. His research interests include robot perception and learning.

Prof. Liu was a recipient of the Andy Chi Best Paper Award in 2017. He has served as the Area Chair for RSS 2018–2019 and a Senior Program Committee Member for IJCAI 2018. He serves as an Associate Editor for some journals, including the IEEE TRANSACTIONS ON AUTOMATION SCIENCE AND ENGINEERING, the IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, the IEEE ROBOTICS AND AUTOMATION LETTERS, *Neurocomputing*, and *Cognitive Computation*, and some conferences, including ICRA and IROS.

Fuchun Sun (Fellow, IEEE) is currently a Full Professor with the Department of Computer Science and Technology, Tsinghua University, Beijing, China. His research interests include intelligent control and robotics.

Prof. Sun was a recipient of the National Science Fund for Distinguished Young Scholars. He serves as the Editor-in-Chief for *Cognitive Computation and Systems* and an Associate Editor for a series of international journals, including the IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS: SYSTEMS and the IEEE TRANSACTIONS ON FUZZY SYSTEMS, MECHATRONICS, AND ROBOTICS AND AUTONOMOUS SYSTEMS.