Pattern Recognition 121 (2022) 108176

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Lifelong robotic visual-tactile perception learning

Jiahua Dong[a,b,c], Yang Cong[a,b,1,*], Gan Sun[a,b], Tao Zhang[a,b,c]

[a] State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang 110016, China
[b] Institutes for Robotics and Intelligent Manufacturing, Chinese Academy of Sciences, Shenyang 110016, China
[c] University of Chinese Academy of Sciences, Beijing 100049, China

ARTICLE INFO

Article history:
Received 27 July 2020
Revised 30 October 2020
Accepted 14 July 2021
Available online 15 July 2021

Keywords:
Lifelong machine learning
Robotics
Visual-tactile perception
Cross-modality learning
Multi-task learning

ABSTRACT

Lifelong machine learning can learn a sequence of consecutive robotic perception tasks via transferring previous experiences. However, 1) most existing lifelong learning based perception methods only take advantage of visual information for robotic tasks, while neglecting another important tactile sensing modality to capture discriminative material properties; 2) Meanwhile, they cannot explore the intrinsic relationships across different modalities and the common characterization among different tasks of each modality, due to the distinct divergence between heterogeneous feature distributions. To address above challenges, we propose a new Lifelong Visual-Tactile Learning (LVTL) model for continuous robotic visual-tactile perception tasks, which fully explores the latent correlations in both intra-modality and cross-modality aspects. Specifically, a modality-specific knowledge library is developed for each modality to explore common intra-modality representations across different tasks, while narrowing intra-modality mapping divergence between semantic and feature spaces via an auto-encoder mechanism. Moreover, a sparse constraint based modality-invariant space is constructed to capture underlying cross-modality correlations and identify the contributions of each modality for new coming visual-tactile tasks. We further propose a modality consistency regularizer to efficiently align the heterogeneous visual and tactile samples, which ensures the semantic consistency between different modality-specific knowledge libraries. After deriving an efficient model optimization strategy, we conduct extensive experiments on several representative datasets to demonstrate the superiority of our LVTL model. Evaluation experiments show that our proposed model significantly outperforms existing state-of-the-art methods with about 1.16%∼15.36% improvement under different lifelong visual-tactile perception scenarios.

1. Introduction

Lifelong machine learning [1,2] has attracted appealing academic interests in the field of robotics, depending on the continuous learning mechanism for a series of different robotic tasks. In the past decades, the intelligent robots are required to equip with human intelligence, *i.e.*, learning and understanding new robotic tasks continuously [3]. Until now, it has been successfully applied into a large number of real-world robotic perception applications, such as SLAM [4], man-machine interaction [5], object classification [6,7], situated robot perception [8], etc.

However, most existing lifelong learning based robotic perception approaches [4,5,7,9] only utilize visual modality information to learn different tasks continuously, while ignoring another important tactile sensing modality (*i.e.*, tactile perception information about hardness, force, temperature and so on). Obviously, tactile perception could efficiently compensate the defects of visual information in many real-world robotic manipulation scenarios [10,11], in which visual information is difficult to distinguish different objects. Take a practical example for demonstration, when a cooking robot grasps vegetables, the visual modality of the vegetables is not informative due to the occlusion of the robot hands. Moreover, some objects with similar appearance and shape (*e.g.*, ripe versus unripe fruits) cannot be discriminated accurately with only visual information. To this end, robots can explore effective information about intrinsic material properties from tactile modality to compensate the deficiency of visual modality by establishing cross-modal correlations. Consequently, the tactile perception modality plays an essential role in supplying complementary information for visual modality to improve the lifelong robotic perception performance.

The straightforward way is to employ current single-view lifelong learning approaches [1,2,12,13] to learn robotic visual-tactile

[*] Corresponding author at: Nanta Street No. 114, Shenyang 110016, China
E-mail addresses: dongjiahua1995@gmail.com (J. Dong), congyang81@gmail.com (Y. Cong), sungan1412@gmail.com (G. Sun), zhangtao2@sia.cn (T. Zhang).
[1] The corresponding author is Prof. Yang Cong.

https://doi.org/10.1016/j.patcog.2021.108176
0031-3203/© 2021 Elsevier Ltd. All rights reserved.

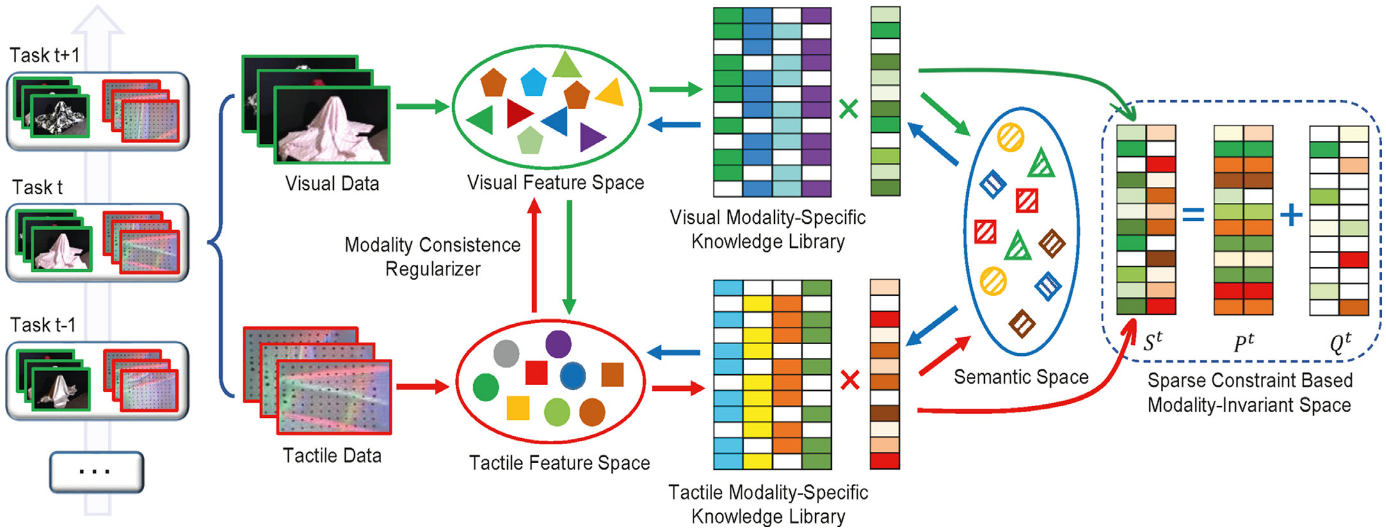

Task t+1
Task t
Task t-1
· · ·
Visual Data
Visual Feature Space
Modality Consistence Regularizer
Tactile Data
Tactile Feature Space
Visual Modality-Specific Knowledge Library
Tactile Modality-Specific Knowledge Library
Semantic Space
S^t
P^t
Q^t
Sparse Constraint Based Modality-Invariant Space

Fig. 1. Illustration of our proposed LVTL model, which consists of *a modality-specific knowledge library* for each modality to explore shared intra-modality information across different tasks, *an auto-encoder mechanism* to mitigate the intra-modality mapping divergence between semantic and feature spaces, *a sparse constraint based modality-invariant space* to capture the intrinsic cross-modality correlations and *a modality* consistency *regularizer* to ensure the semantic consistency across visual and tactile modalities.

tasks continuously, by integrating both visual and tactile modalities into a high dimensional feature space. Nevertheless, it can inevitably result in high computational costs and storage (*e.g.*, large infrastructures), which are not satisfied in real-world robotic applications. Besides, these methods treat both visual and tactile modalities independently and equally with one common knowledge library. They cannot efficiently explore the intrinsic correlations in cross-modality while highlighting the common characterization across different tasks in intra-modality, since the distribution gap between visual and tactile modalities is apparently large [14]. For example, the sensor devices and information format are different across visual and tactile modalities. Specifically, visual sensor normally captures color, global shape and rough texture information, but touch sensor perceives detailed texture, hardness and temperature via constant physical contact. Therefore, in this paper, how to explore the intrinsic complementary knowledge across visual and tactile modalities under the lifelong learning manner is our main focus.

To tackle above challenges, as depicted in Fig. 1, a new Lifelong Visual-Tactile Learning (*i.e.*, LVTL) model is developed, which is equipped with powerful continuous learning ability for new robotic visual-tactile tasks. Our proposed LVTL model could efficiently capture the common characterizations across different tasks in intra-modality aspect, while exploring the latent complementary correlations in cross-modality aspect. To be specific, for intra-modality correlations, we construct a modality-specific knowledge library for each modality to capture the shared experience knowledge across different tasks. Particularly, it encourages the learned and new coming tasks to share a common embedding space in intra-modality, while ensuring the consistent intra-modality mapping between semantic and feature spaces by incorporating with an auto-encoder mechanism. Furthermore, for cross-modality relationships, we construct a sparse constraint based modality-invariant space to explore the shared complementary knowledge across visual and tactile modalities, and quantify the importance of each modality for the consecutive visual-tactile robotic perception tasks at the same time. Besides, a modality consistency regularizer is employed to further mitigate the heterogeneous distribution gap among visual and tactile modalities, while minimizing semantic divergence among different modality-specific knowledge libraries. Due to the non-convex and NP-hard formulation, we utilize the alternating direction strategy to optimize our LVTL model when new visual-tactile perception tasks arrive continuously. To the end, comprehensive experiments on several visual-tactile benchmark datasets demonstrate the effectiveness and efficiency of our LVTL model. Particularly, when compared with existing state-of-the-art methods, our proposed LVTL model achieves significant performance improvement (*i.e.*, 1.44%∼10.18% and 1.16%∼15.36% in terms of all evaluation metrics) on PHAC-2 and GelFoldFabric datasets. Moreover, although there are a large number of consecutive robotic visual-tactile tasks, our LVTL model could consume less computational time (*i.e.*, 6.43 s and 19.56 s) on PHAC-2 dataset with 27 tasks and GelFoldFabric dataset with 60 tasks.

The main contributions of this paper are summarized as follows:

- We develop a new Lifelong Visual-Tactile Learning (LVTL) model to learn a sequence of robotic visual-tactile perception tasks continuously. To the best of our knowledge, this is an earlier exploration about robotic visual-tactile cross-modality learning under the lifelong learning manner.
- We design a modality-specific knowledge library for each modality to capture common intra-modality knowledge across different tasks, by preserving the shared experience information of the learned and new coming robotic visual-tactile tasks.
- A sparse constraint based modality-invariant space is constructed to explore the shared complementary knowledge across visual and tactile modalities, and identify the importance of each modality for the new coming robotic visual-tactile tasks simultaneously.

2. Related work

We introduce some representative researches about visual-tactile perception and lifelong machine learning in this section.

2.1. Visual-tactile perception

In the field of robotics, vision and touch [14] are both the important modality perception information, which are successfully applied into a large number of real-world robotic tasks. Generally, these robotic application scenarios accessing visual and tactile modalities information are mainly divided into three categories, *i.e.*, 3-D object reconstruction [16,17], object recognition [10,11,18] and cross-modality consistency matching [19,20,45].

Specifically, for 3-D object reconstruction tasks, Illonen et al. [16] propose a complete 3-D reconstruction method for unknown objects by incorporating the fused information of visual and tactile modalities. Wang et al. [17] first employ shape priors and vision perception to perform 3D object shape prediction from a single-view color image, and then refines the predicted shape with tactile information. Besides, motivated by the kernel sparse coding strategy, Liu et al. [10] aim to recognize household objects by developing a novel visual-tactile fusion network. Yuan et al. [18] focus on constructing a robot system for material property recognition of the clothing according to the tactile modality information, where the robot system is composed of a high-resolution tactile sensor and a monocular camera. Furthermore, Yuan et al. [11] explore the intrinsic correlations between visual and tactile modalities, by capturing the underlying properties of fabrics via a multi-input network. As for cross-modality consistency matching, Li et al. [19] utilize conditional generative adversarial network to conduct visual and tactile consistency measurement. For visual and tactile perception, Lee et al. [20] generate pseudo visual images relying on the data from tactile modality, and vice verse. Moreover, Zhang et al. [45] employ deep non-negative matrix factorization technology to explore the visual-tactile clustering consistency. Generally, visual-tactile perception information plays an important role in robotic applications. However, these existing approaches assume that all training data is available at once instead of the lifelong learning settings, which results in high computational costs and storage (*e.g.*, large infrastructures).

2.2. Lifelong machine learning

The lifelong learning methods aim to transfer learned knowledge from previous tasks to the new ones [21]. Based on the existing multi-task learning formulation [22], an efficient lifelong learning model (*i.e.*, ELLA) [1] is developed to learn a series of tasks continuously. To be specific, ELLA [1] constructs a shared common knowledge library to preserve the learning experience of multiple related tasks. It could learn new observed tasks consecutively by constructing knowledge library to transfer effective learning experience. Isele et al. [12] propose to explore high-level task descriptors for zero-shot lifelong learning, which could effectively model the inter-task correlations by incorporating a coupled dictionary. Motivated by Ruvolo and Eaton [1] and Isele et al. [12], Sun et al. [2] develop an active lifelong learning model to automatically identify new or learned tasks via a “Watchdog” mechanism. Different from them, [23] focuses on learning an inductive bias in form of a transfer procedure for continuous tasks with non-stationary distributions. However, above single-view lifelong learning models cannot be directly applied into more challenging real-world scenarios with multiple data modalities, such as smart grid field [24] and attributed graph clustering system [25], since they fail to capture shared knowledge and discriminative complementary information across different modalities. Therefore, for the consecutive learning tasks with multiple views [26], Li et al. first propose a decision function in the shared latent space to integrate lifelong learning into multi-view learning [50]. Sun et al. [27] employ the robust multi-task learning formulation to explore the shared representations across different views. Sun et al. [7] utilize deep matrix factorization to learn multi-view tasks continuously.

To extend lifelong learning strategy into robotic applications, Thrun et al. [3] focus on transferring knowledge between different robotic tasks . It deals with the scenarios where robots encounter a large number of learning tasks over their lifespans. Glover and Wyeth [5] develop computational models of affordance to achieve objects interaction under the lifelong learning manner. Young et al. [8], Hawasly and Ramamoorthy [29] integrate robot perception and semantic web mining for lifelong object learning. Xia et al. [9] aim to address large-scale social media sentiment analysis by introducing distantly supervised lifelong learning. Zheng et al. [30] develop a convolution neural network based lifelong learning framework for robotic visual-tactile cross-modal material perception. Liu et al. [31] employ an online dictionary learning algorithm to learn the consecutive robotic tasks with multiple modalities (*i.e.*, visual, tactile and auditory modalities).

According to above discussions, the most relevant work to our LVTL model is [31]. However, there are still some significant differences between our LVTL model and [31]: 1) Our model constructs a modality-specific knowledge library for each modality to explore distinctive intra-modality representations, due to the heterogeneous feature distributions across modalities. Liu et al. [31] learn a shared knowledge dictionary for all modalities, which cannot efficiently capture the modality-specific knowledge. 2) Our model proposes a sparse constraint based modality-invariant space, which could explore the complementary knowledge across modalities and quantify the contributions of each modality for lifelong robotic learning. However, [31] cannot achieve it via a common knowledge dictionary of all modalities. 3) Compared with the knowledge library in [31] that is constrained by Frobenius norm, we propose a modality consistency regularizer and an auto-encoder mechanism to construct more discriminative modality-specific knowledge libraries.

3. The proposed model

3.1. Problem statement

Suppose that a lifelong robotic visual-tactile perception system faces a series of consecutive learning tasks: $\{\mathcal{Z}^t\}_{t=1}^T$, where T is the total number of learning tasks. In the tth task $\mathcal{Z}^t = \left\{\{f_m^t, X_m^t\}_{m=1}^M, Y^t\right\}$, $X_m^t = [x_{1m}^t, x_{2m}^t, \cdots, x_{n_t m}^t] \in \mathbb{R}^{d_m \times n_t}$ denotes n_t data samples of the mth modality represented by d_m dimensional features, and $Y^t = [y_1^t, y_2^t, \cdots, y_{n_t}^t] \in \mathbb{R}^{n_t}$ represents the corresponding labels. M is the number of modalities, and we set $M = 2$ (*i.e.*, visual and tactile modalities) in this paper. To be specific, $m = 1$ denotes the visual modality and $m = 2$ represents the tactile modality. We aim to learn a linear mapping $f_m^t : X_m^t \rightarrow Y^t$ for the mth modality in the tth task, and the mapping f_m^t can be expressed as $Y^t = f_m^t(X_m^t; w_m^t) = \langle X_m^t, w_m^t \rangle$, where $w_m^t \in \mathbb{R}^{d_m}$ denotes the corresponding task classifier. When the data from the tth learning task arrives, the lifelong visual-tactile perception system needs to make the predictions for both previously learned $t-1$ tasks and the current tth task. Note that the lifelong robotic visual-tactile perception system has no any prior knowledge about the total number of tasks and each task distribution.

3.2. Lifelong learning for visual modality

Generally, most existing lifelong learning models [1,2] are designed to learn consecutive single-view visual tasks, which neglect the tactile perception modality that could provide complementary material property information for robotic tasks. Specifically, they assume that the learned task classifier w_m^t of visual modality for each task shares a common knowledge library $L \in \mathbb{R}^{d_m \times k}$, where k is the number of basic tasks. In other words, the task classifier w_m^t can be formulated as $w_m^t = L s_m^t$, where $s_m^t \in \mathbb{R}^k$ denotes the sparse coding [32,51] over the knowledge library L that selects few atoms to construct the task classifier w_m^t. For $\{w_m^t, s_m^t\}$ in this subsection, we set $m = 1$ for visual modality. Based on above assumption, the general lifelong learning for visual modality tasks $(m = 1)$ can be expressed as:

$$\min_{L} : \frac{1}{T}\sum_{t=1}^{T}\min_{s_m^t} : \left\{\frac{1}{n_t}\sum_{i=1}^{n_t}\mathcal{L}(f_m^t(x_{im}^t, Ls_m^t), y_i^t) + \mu_1\|s_m^t\|_1\right\} + \mu_2\|L\|_F^2, \tag{1}$$

where $\mathcal{L}(\cdot)$ represents the loss function (such as squared loss, logistic loss, etc). $\mu_1 > 0$ and $\mu_2 > 0$ are the balanced parameters.

Obviously, the formulation in Eq. (1) could efficiently make predictions for a series of consecutive tasks with only visual modality by learning a shared knowledge library L. However, it cannot be directly applied into the lifelong robotic visual-tactile system, since it treats all modalities independently and equally by merging visual and tactile samples into one high dimensional feature to learn a shared knowledge library. Moreover, it cannot explore the intrinsic intra-modality and cross-modality relationships, due to the heterogeneous distribution divergence across different visual-tactile tasks and modalities.

3.3. Lifelong visual-tactile learning (LVTL)

In this subsection, we detailedly introduce our proposed LVTL model to address above challenges, as shown in Fig. 1. Our LVTL model could efficiently capture the intra-modality correlations across different visual-tactile tasks and the complementary cross-modality relationships, which are presented as follows:

Intra-Modality Correlations: Define $\mathcal{M} = \{1, \cdots, M\}$ as a finite set of available modalities. We set $M = 2$ in the lifelong robotic visual-tactile system, *i.e.*, visual and tactile modalities. In the tth learning task, $x_{im}^t \in \mathbb{R}^{d_m} (m \in \mathcal{M})$ denotes d_m dimensional features of the ith data sample for the mth modality. To explore the intrinsic intra-modality relationships across different tasks, we construct a shared modality-specific knowledge library for each modality via matrix factorization technique, as shown in Fig. 1. The different task classifiers $\{w_m^t\}_{t=1}^T$ in the mth modality can be described as the linear combination of bases in modality-specific knowledge library L_m. Consequently, for the tth task, the task classifier w_m^t of the mth modality (vision or touch) is formulated as:

$$w_m^t = L_m s_m^t, \quad \forall m \in \mathcal{M} = \{1, \cdots, M\}, \tag{2}$$

where $L_m \in \mathbb{R}^{d_m \times k}$ is the shared modality-specific knowledge library across different tasks in the mth modality. $s_m^t \in \mathbb{R}^k$ denotes the latent representation for the tth task in the mth modality. k is the number of representative knowledge bases in L_m. Obviously, the discriminative relationships across a series of consecutive visual-tactile tasks can be efficiently characterized via the shared modality-specific knowledge libraries $\{L_1, \cdots, L_M\}$.

Furthermore, an auto-encoder mechanism is developed to further promote the robustness and generalization of modality-specific knowledge library for different robotic tasks. Specifically, it efficiently narrows the intra-modality mapping divergence between feature and semantic spaces, while ensuring their consistency across different visual-tactile tasks. Concretely, the modality-specific knowledge library L_m for the mth modality can be optimized by:

$$\min_{L_m} : \frac{1}{T}\sum_{t=1}^{T}\left\{\frac{1}{n_t}\sum_{i=1}^{n_t}\mathcal{L}\Big(f_m^t(x_{im}^t, L_m s_m^t), y_i^t\Big) + \mathcal{L}\Big(f_m^t((L_m s_m^t)^\top, (y_i^t)^\top), x_{im}^t\Big)\right\}, \tag{3}$$

where $\mathcal{L}(\cdot)$ represents the loss function, which is defined in Eq. (1). The encoder $f_m^t(x_{im}^t, L_m s_m^t)$ is defined as $f_m^t(x_{im}^t, L_m s_m^t) = (x_{im}^t)^\top L_m s_m^t$, and shares the similar definition with the decoder $f_m^t((L_m s_m^t)^\top, (y_i^t)^\top)$.

Cross-Modality Relationships: Although the formulation in Eq. (3) could explore the intra-modality intrinsic relationships across different visual-tactile tasks, it cannot fully leverage the cross-modality complementary information, which significantly degrades the lifelong learning performance. To this end, we propose to establish a sparse constraint based modality-invariant space, which reconstructs the latent sparse representation for each task by two collaborative components. Specifically, the sparse coding optimization of the mth modality (*i.e.*, s_m^t) is considered as an individual task, and our model focuses on learning the latent representation $S^t = \{s_1^t, \cdots, s_M^t\} \in \mathbb{R}^{k \times M}$ for the tth task. To capture complementary cross-modality correlations, as depicted in Fig. 1, the latent representation S^t can be reconstructed by two collaborative components P^t and Q^t, *i.e.*, $S^t = P^t + Q^t$. Concretely, $P^t \in \mathbb{R}^{k \times M}$ aims to explore shared knowledge across different modalities via a row sparse constraint, while $Q^t \in \mathbb{R}^{k \times M}$ identifies the contributions of different modalities for the tth task via a column sparse constraint. The latent representation S^t for the tth task can be obtained by optimizing the following objective:

$$\begin{aligned}\min_{S^t} : \quad & \tfrac{1}{n_t}\sum_{i=1}^{n_t}\tfrac{1}{M}\sum_{m=1}^{M}\mathcal{L}\Big(f_m^t(x_{im}^t, L_m s_m^t), y_i^t\Big) + \mathcal{L}\Big(f_m^t((L_m s_m^t)^\top, (y_i^t)^\top), x_{im}^t\Big) \\ & +\lambda_1\|P^t\|_{1,\infty} + \lambda_2\|(Q^t)^\top\|_{1,\infty}, \\ & s.t. \quad S^t = P^t + Q^t,\end{aligned} \tag{4}$$

where $\lambda_1 > 0$ and $\lambda_2 > 0$ are the balanced parameters. Intuitively, $\|P^t\|_{1,\infty}$ encourages different modality-specific knowledge libraries to be more consistent for the shared complementary information across modalities, by capturing the shared atoms among different libraries. Meanwhile, $\|(Q^t)^\top\|_{1,\infty}$ quantifies the importance of different modalities for the tth task. When the mth column of Q^t is greater than zero, the knowledge library L_m of the mth modality play a more essential role in the tth task, and vice verse.

Given a sequence of consecutive robotic visual-tactile tasks, the overall optimization objective that incorporates both intra-modality and cross-modality correlations is formally formulated as follows:

$$\begin{aligned}\min\nolimits_{L_m} : \tfrac{1}{T}\sum_{t=1}^{T}\min_{S^t} : \Big\{ & \tfrac{1}{n_t}\sum_{i=1}^{n_t}\tfrac{1}{M}\sum_{m=1}^{M}\mathcal{L}\Big(f_m^t(x_{im}^t, L_m s_m^t), y_i^t\Big) \\ & +\mathcal{L}\Big(f_m^t((L_m s_m^t)^\top, (y_i^t)^\top), x_{im}^t\Big) + \lambda_1\|P^t\|_{1,\infty} + \lambda_2\|(Q^t)^\top\|_{1,\infty}\Big\} \\ & + \sum_{m\neq n}^{M}\gamma_m\|(X_m^t)^\top L_m - (X_n^t)^\top L_n\|_F^2, \\ & s.t. \quad S^t = P^t + Q^t, \quad \forall t = 1, \cdots, T\end{aligned} \tag{5}$$

where $\{\gamma_m\}_{m=1}^M$ are the balanced parameters. X_m^t and X_n^t denote the visual and tactile data in the tth task, respectively. The modality consistency regularizer $\|(X_m^t)^\top L_m - (X_n^t)^\top L_n\|_F^2$ is designed to ensure the semantic consistency between different modality-specific libraries, while aligning the heterogeneous distribution gap across different modalities. Intuitively, it encourages the samples from visual and tactile modalities to share the common knowledge bases for different modality-specific libraries.

4. Model optimization

In this section, the optimization procedure for solving Eq. (5) is elaborated. Since the formulation in Eq. (5) is not convex with respect to L_m and S^t jointly, it is non-trivial to solve this optimization problem. To this end, we first introduce how to approximate Eq. (5) using the Taylor expansion [1], and then present an alternating optimization strategy to update L_m and S^t iteratively.

4.1. Taylor expansion

For simplification, we define the first two terms in Eq. (5) as $\mathcal{L}_m^t(L_m s_m^t)$, *i.e.*,

$$\mathcal{L}_m^t(L_m s_m^t) = \frac{1}{n_t}\sum_{i=1}^{n_t} \mathcal{L}\big(f_m^t(x_{im}^t, L_m s_m^t), y_i^t\big) + \mathcal{L}\big(f_m^t((L_m s_m^t)^\top, (y_i^t)^\top), x_{im}^t\big). \tag{6}$$

Motivated by the Taylor expansion [1], for the m-th modality, we utilize the second-order Taylor expansion of $\mathcal{L}_m^t(L_m s_m^t)$ around $L_m s_m^t = w_m^t$ to approximate the first two terms in Eq. (5), which is presented as follows:

$$\begin{aligned}\mathcal{L}_m^t(L_m s_m^t) &= \mathcal{L}_m^t(w_m^t) + \langle \nabla w_t^m \mathcal{L}(L_m^t s_m^t), L_m s_m^t - w_m^t\rangle \\ &\quad + \frac{1}{2}\|L_m s_m^t - w_m^t\|^2_{\mathcal{H}_m^t},\end{aligned} \tag{7}$$

where $\nabla w_t^m \mathcal{L}_m^t(L_m^t s_m^t)$ denotes the first-order gradient information with respect to w_m^t, and $\mathcal{H}_m^t$ represents the Hessian matrix of $\mathcal{L}_m^t(L_m^t s_m^t)$ around w_m^t. Generally, for the mth modality in the tth task, w_m^t and $\mathcal{H}_m^t$ in Eq. (7) are defined as follows:

$$\begin{aligned} w_m^t &= \arg\min_{L_m s_m^t} \frac{1}{n_t}\sum_{i=1}^{n_t} \mathcal{L}\big(f_m^t(x_{im}^t, L_m s_m^t), y_i^t\big) \\ &\quad + \mathcal{L}\big(f_m^t((L_m s_m^t)^\top, (y_i^t)^\top), x_{im}^t\big), \\ \mathcal{H}_m^t &= \nabla^2_{L_m s_m^t, L_m s_m^t} \mathcal{L}_m^t(L_m^t s_m^t)|_{L_m s_m^t = w_m^t}. \end{aligned} \tag{8}$$

Note that we can easily compute w_m^t and $\mathcal{H}_m^t$ for the mth modality via Eq. (8), when the robotic visual-tactile data from the tth task arrives.

After plugging the second-order Taylor expansion of $\mathcal{L}_m^t(L_m^t s_m^t)$ into Eq. (5) and substituting S^t as $P^t + Q^t$, we can obtain the following optimization objective:

$$\begin{aligned} \min_{L_m} : \frac{1}{T}\sum_{t=1}^{T} \min_{P^t, Q^t} : \Bigg\{ \frac{1}{M}\sum_{m=1}^{M} \|w_m^t - L_m(p_m^t + q_m^t)\|^2_{\mathcal{H}_m^t} \\ + \lambda_1\|P^t\|_{1,\infty} + \lambda_2\|(Q^t)^\top\|_{1,\infty} \Bigg\} \\ + \sum_{m \neq n}^{M} \gamma_m \|(X_m^t)^\top L_m - (X_n^t)^\top L_n\|_F^2, \end{aligned} \tag{9}$$

where p_m^t and q_m^t represent the mth column of P^t and Q^t, respectively. After computing w_m^t and $\mathcal{H}_m^t$ for the tth task via Eq. (8), we can update L_m and $\{P^t, Q^t\}$ in Eq. (9) iteratively via an alternating optimization strategy.

4.2. Updating $\{P^t, Q^t\}$ when fixing L_m

When the modality-specific library L_m is fixed, the formulation in Eq. (9) for optimizing variables $\{P^t, Q^t\}$ can be rewritten as follows:

$$\min_{P^t, Q^t} : \frac{1}{M}\sum_{m=1}^{M} \|w_m^t - L_m(p_m^t + q_m^t)\|^2_{\mathcal{H}_m^t} + \lambda_1\|P^t\|_{1,\infty} + \lambda_2\|(Q^t)^\top\|_{1,\infty}. \tag{10}$$

Due to the non-convex property of $\|\cdot\|_{1,\infty}$-norm constraint, we employ Proximal Alternating Linearized Minimization (PALM) [33] to solve variables $\{P^t, Q^t\}$ in Eq. (10), which is formulated as:

$$P^t, Q^t = \arg\min_{P, Q} : g(P, Q) + r_1(P) + r_2(Q), \tag{11}$$

where $g(P, Q) = \frac{1}{M}\sum_{m=1}^{M} \|\sqrt{\mathcal{H}_m^t} w_m^t - \sqrt{\mathcal{H}_m^t} L_m(p_m^t + q_m^t)\|_F^2$ is a convex function when both w_m^t and L_m are multiplied by $\sqrt{\mathcal{H}_m^t}$. Moreover, $r_1(P) = \lambda_1\|P\|_{1,\infty}$ and $r_2(Q) = \lambda_2\|Q^\top\|_{1,\infty}$ represent the regularizers. As the convex function $g(U)$ at the previous approximate solution U_{l-1} could be regularized by the quadratic proximal term, the current approximate solution U_l could be solved by:

$$U_l = \arg\min_{U} : g(U_{l-1}) + \langle \nabla_U g(U_{l-1}), U - U_{l-1}\rangle + \frac{\xi_l}{2}\|U - U_{l-1}\|_F^2, \tag{12}$$

where $U = \binom{P}{Q}$, and $\nabla_U g(U_{l-1})$ represents the first-order gradient of $g(U_{l-1})$. The stepsize parameter ξ_l is appropriately determined by the backtracking rule [34]. We then add the regularization terms $r_1(P)$ and $r_2(Q)$ to Eq. (12), and remove the irrelevant constant terms. The optimization subproblems to solve P and Q are introduced as follows:

$$\begin{aligned} P_l &= \arg\min_{P} : \tfrac{\xi_l}{2}\|P - (P_{l-1} - \tfrac{1}{\xi_l}\nabla_P g(U_{l-1}))\|_F^2 + \lambda_1\|P\|_{1,\infty}, \\ Q_l &= \arg\min_{Q} : \tfrac{\xi_l}{2}\|Q - (Q_{l-1} - \tfrac{1}{\xi_l}\nabla_Q g(U_{l-1}))\|_F^2 + \lambda_2\|Q^\top\|_{1,\infty}. \end{aligned} \tag{13}$$

Motivated by Liu and Ye [35], Eq. (13) can be solved efficiently by updating each row of P and $Q^\top$ separately, due to the separable structure involved in Eq. (13). When the convergence condition of solving Eq. (13) is satisfied, we can obtain the final solution $\{P^t, Q^t\}$ for Eq. (11). Moreover, the optimization strategy for updating $\{P^t, Q^t\}$ is summarized in Algorithm 1.

Algorithm 1 Updating $\{P^t, Q^t\}$ via PALM Strategy [33].

```
Input: ${w_m^t, \mathcal{H}_m^t, L_m}_{m=1}^M$, $\lambda_1 > 0, \lambda_2 > 0$ and MAX-ITER;
Output: ${P^t, Q^t}$;
Initialize: $P_0 \in \mathbb{R}^{k\times M}, Q_0 \in \mathbb{R}^{k\times M}, \xi_0 > 0$;
for $l = 1, \ldots,$ MAX-ITER do
    Solve ${P_l, Q_l}$ via Eq. (13);
    Update $\xi_l$ via backtracking rule [34];
    if convergence condition is satisfied then
        $P^t = P_l, Q^t = Q_l$;
        break;
    end if
end for
Return ${P^t, Q^t}$;
```

4.3. Updating L_m when fixing $\{P^t, Q^t\}$

When the variables $\{P^t, Q^t\}$ are solved via Eq. (11), the optimization for modality-specific knowledge library L_m in Eq. (9) has closed solution. To update L_m, we substitute $p_m^t + q_m^t$ as s_m^t, and obtain the following formulation:

$$\begin{aligned} \min_{L_m} : \frac{1}{T}\sum_{t=1}^{T} \frac{1}{M}\sum_{m=1}^{M} \|w_m^t - L_m(p_m^t + q_m^t)\|^2_{\mathcal{H}_m^t} \\ + \sum_{m \neq n}^{M} \gamma_m \|(X_m^t)^\top L_m - (X_n^t)^\top L_n\|_F^2. \end{aligned} \tag{14}$$

After equating the gradient of Eq. (14) to zero, we can update the column-wise vectorization of L_m via $(R_m)^{-1}V_m$ (*i.e.*, $\mathrm{vec}(\mathrm{L_m}) = (\mathrm{R_m})^{-1}\mathrm{V_m}$), where $\mathrm{vec}(\cdot)$ represents the column-wise vectorization operation, $R_m \in \mathbb{R}^{(kd_m)\times(kd_m)}$ and $V_m \in \mathbb{R}^{kd_m}$ denotes the statistical records of the mth modality. To store previous knowledge of each modality, R_m and V_m can be updated by the following strategies:

$$\begin{aligned} R_m &= \tfrac{1}{T}\sum_{t=1}^{T}\left\{\left(s_m^t(s_m^t)^\top\right)\otimes\mathcal{H}_m^t + \gamma_m\sum_{m\neq n}^{M}\gamma_m I_k\otimes\left(X_m^t(X_m^t)^\top\right)\right\},\\ V_m &= \tfrac{1}{T}\sum_{t=1}^{T}\left\{\left(s_m^t\otimes(\mathcal{H}_m^t w_m^t)\right)+\mathrm{vec}\left(L_n^\top X_n^t(X_m^t)^\top\right)\right\}, \end{aligned} \tag{15}$$

where $\otimes$ denotes the Kronecker product. R_m and V_m are incremental continuously as the new tasks coming. Moreover, the whole optimization strategy of our model is presented in Algorithm 2.

Algorithm 2 The Optimization Procedure of Our LVTL Model.

```
Input: A series of visual-tactile tasks $\left\{\{X_m^t\}_{m=1}^M, Y^t\right\}_{t=1}^T$, $\gamma_m > 0$,
    $R_m = \mathbf{0}_{kd_m,kd_m}$, $V_m = \mathbf{0}_{kd_m,1}$;
Output: $\{L_m\}_{m=1}^M$, $\{S^t\}_{t=1}^T$;
Initialize: $\{L_m\}_{m=1}^M$;
for $t = 1, \dots, T$ do
    The $t$th new task $\{X^t, Y^t\}$ arrives;
    for $m = 1, \dots, M$ do
        Compute $\{w_m^t, \mathcal{H}_m^t\}$ via Eq. (8);
        Compute $S^t$ via Algorithm 1;
        Update $R_m$: $R_m \leftarrow R_m + \Big(\big(s_m^t(s_m^t)^\top\big)\otimes\mathcal{H}_m^t + \gamma_m\sum_{m\neq n}^M \gamma_m I_k\otimes\big(X_m^t(X_m^t)^\top\big)\Big)$;
        Update $V_m$: $V_m \leftarrow V_m + \Big(\big(s_m^t\otimes(\mathcal{H}_m^t w_m^t)\big) + \mathrm{vec}\big(L_n^\top X_n^t(X_m^t)^\top\big)\Big)$;
        Update $L_m$: $L_m \leftarrow (\frac{1}{t}R_m)^{-1}(\frac{1}{t}V_m)$;
    end for
end for
Return $\{L_m\}_{m=1}^M$, $\{S^t\}_{t=1}^T$;
```

4.4. Computational complexity

The main computational complexity of optimizing our LVTL model via Algorithm 2 involves three subproblems, which are presented as follows: 1) The complexity of computing $\{w_m^t, \mathcal{H}_m^t\}$ is $O(\sum_{m=1}^M \eta(d_m, n_t))$, where $\eta(\cdot)$ represents the complexity of each task learner [1]. 2) Optimizing $S^t = P^t + Q^t$ via Eq. (11) costs $O\big(\sum_{m=1}^M((d_m)^3 + k(d_m)^2 + kd_m) + k\mathrm{log}k + M\mathrm{log}M\big)$. Specifically, multiplying w_m^t and L_m by $\sqrt{\mathcal{H}_m^t}$ consumes $O\big(\sum_{m=1}^M(d_m)^3 + k(d_m)^2\big)$. The gradient computation for S^t is $O(\sum_{m=1}^M kd_m)$, and the optimization complexity for Eq. (13) is $O(k\mathrm{log}k + M\mathrm{log}M)$. 3) Updating L_m involves the computation of R_m and V_m, which consumes $O(\sum_{m=1}^M k^2(d_m)^3)$. Overall, when the new visual-tactile task comes, the total computational complexity of our LVTL model is $O\big(\sum_{m=1}^M(\eta(d_m, n_t) + (d_m)^3 + k(d_m)^2 + k^2(d_m)^3)\big)$. Note that $M = 2$ (*i.e.*, visual and tactile modalities), and we set k as 2 or 3 in our experiments. When compared with the feature dimension d_m of each modality and the number of samples n_t, the values of M and k are usually small. Therefore, our proposed LVTL model is computationally efficient under the lifelong visual-tactile perception scenarios.

5. Experiments

We report extensive comparison experiments between our model and some representative models to illustrate the effectiveness of our model in this section.

5.1. Datasets and evaluation metric

Generally, three benchmark datasets are utilized to conduct empirical comparisons in our experiments:

PHAC-2 [2]: This visual-tactile perception dataset is collected from 53 different objects, and each object is composed of 10 tactile signals and 8 color images. The first 8 tactile signals and all images are used for comparison experiments in this paper. We employ generative adversarial network [36,52] to achieve feature augmentation, which generates 60 valid samples with visual and tactile modalities for each object. Inspired by Gao et al. [37], we extract 1024-D and 512-D features for visual and tactile modalities via VGG-16 framework [38], respectively. Each learning task consists of two different classes, and there are no same class between any two tasks. The number of total robotic learning tasks with random task order is 27.

GelFoldFabric [3]: This visual-tactile perception dataset consists of 118 kinds of fabrics, where each category contains 10 color images and 10 tactile images. After utilizing generative adversarial network [36,52] to generate 60 training samples for each category via feature augmentation, we employ pre-trained VGG-16 network [38] to extract 1024-D visual features and 512-D tactile features. We consider two different fabric categories for each learning task, and any two tasks cannot contain the same fabric category. We set the total robotic visual-tactile tasks with random task order as 60.

SVHN [4]: SVHN [39] with 10 different categories is a real-world large-scale challenging object recognition dataset, which is employed to validate the generalization performance and scalable application effectiveness of our proposed LVTL model. It consists of over 70,000 samples that are collected from the street view house numbers. For this dataset, the RGB images and their corresponding gray samples are regarded as different modalities. We utilize the pre-trained VGG-16 network [38] to respectively extract 1024-D and 512-D features for RGB and gray modalities. We consider each object category as one learning task, and set the positive and negative patterns with the same samples number in each task. The number of learning tasks for SVHN dataset is 10, and all tasks are sorted randomly.

Experimental Configurations: For each evaluation dataset, we randomly select half of the available samples for training and the other half for evaluation. Moreover, each task in each dataset is presented consecutively to our proposed LVTL model, which satisfies the lifelong robotic learning settings. In our experiments, all reported results are averaged over 5 random runs.

Evaluation Metric: To justify the superiority of our model, five different metrics, *i.e.*, Area Under Curve (Auc), Macro-F1, Micro-F1, Accuracy (Acc) and Recall, are employed for performance evaluation.

5.2. Competing methods

To demonstrate the effectiveness of our LVTL model, we introduce comparison experiments with several state-of-the-art methods, including two multi-task learning models (*i.e.*, MTFL [40] and rTGRT [41]), two multi-view multi-task approaches (*i.e.*, MAMUDA [42] and lslMTMV [43]), and four lifelong learning methods (*i.e.*, ELLA [1], lslMTMV [43], rLM^{2}L [27] and L^{2}HMT [31]). Generally, these competing models are detailedly introduced as follows:

MTFL [40] and **rTGRT** [41] focus on exploring a common set of features among relevant tasks, while identifying outlier tasks by concatenating the features from different modalities of each task into a high dimensional feature. They assume that the training data of all robotic learning tasks are available at once instead of the lifelong learning manner. Similar to [40,41], **MAMUDA** [42] and **lslMTMV** [43] have access to all learning tasks at once as well.

[2] http://people.eecs.berkeley.edu/yg/icra2016

[3] http://people.csail.mit.edu/yuanwz/fabric-perception.htm

[4] http://ufldl.stanford.edu/housenumbers

Table 1
Performance comparisons (%) between our LVTL model and state-of-the-art methods in terms of five metrics (mean ± standard deviation) on PHAC-2 dataset.

Metric	MTFL [40]	rTGRT [41]	MAMUDA [42]	MFM [44]	ELLA [1]	lslMTMV [43]	rLM^{2}L [27]	L^{2}HMT [31]	Ours
Auc	89.19±0.20	89.32±0.08	88.16±0.49	89.70±0.82	87.50±0.18	91.50±0.18	90.86±0.55	89.13±0.24	**92.94±0.49**
Macro-F1	87.17±0.31	87.96±0.15	86.32±0.39	89.28±0.74	90.91±0.36	89.05±0.36	90.54±0.62	88.41±0.34	**93.58±0.32**
Micro-F1	84.45±0.16	85.47±0.63	85.16±0.72	87.90±0.69	90.89±0.42	85.67±0.41	88.26±0.28	86.19±0.30	**92.81±0.27**
Acc	81.02±0.74	82.39±0.21	84.80±0.66	86.38±0.93	87.50±0.73	83.33±0.54	86.11±0.65	84.74±0.53	**91.20±0.58**
Recall	86.79±1.05	85.62±0.85	86.42±1.13	88.33±0.51	90.39±0.20	87.53±0.38	90.37±0.73	90.04±0.80	**93.46±0.66**

Table 2
Performance comparisons (%) between our LVTL model and state-of-the-art methods in terms of five metrics (mean ± standard deviation) on GelFoldFabric dataset.

Metric	MTFL [40]	rTGRT [41]	MAMUDA [42]	MFM [44]	ELLA [1]	lslMTMV [43]	rLM^{2}L [27]	L^{2}HMT [31]	Ours
Auc	92.20±0.24	94.01±0.53	90.47±0.31	91.53±0.43	95.92±0.26	95.02±0.27	95.12±0.24	94.73±0.41	**97.21±0.18**
Macro-F1	86.01±0.31	88.54±0.44	85.73±0.19	86.46±0.27	84.21±0.74	88.56±0.41	92.33±0.31	91.96±0.18	**94.87±0.25**
Micro-F1	79.83±0.28	82.36±0.47	85.42±0.21	84.63±0.52	84.23±0.55	84.36±0.31	91.58±0.16	91.37±0.24	**94.25±0.32**
Acc	79.33±0.51	82.17±0.22	81.62±0.85	83.04±0.64	78.57±1.04	85.48±0.19	91.35±0.27	90.95±0.53	**93.93±0.35**
Recall	83.06±0.20	84.87±0.39	80.71±0.38	82.78±0.86	93.42±1.32	85.21±0.52	92.08±0.43	91.44±0.47	**94.58±0.21**

Table 3
Performance comparisons (%) between our LVTL model and state-of-the-art methods in terms of five metrics (mean ± standard deviation) on SVHN dataset.

Metric	MTFL [40]	rTGRT [41]	MAMUDA [42]	MFM [44]	ELLA [1]	lslMTMV [43]	rLM^{2}L [27]	L^{2}HMT [31]	Ours
Auc	86.96±0.41	88.31±0.60	87.58±0.52	88.61±0.34	97.51±0.16	69.37±0.46	96.19±0.72	95.03±0.27	**99.49±0.13**
Macro-F1	87.15±0.83	89.02±0.29	85.23±0.17	83.37±0.19	91.77±0.48	53.60±0.29	91.12±0.49	89.95±0.16	**94.97±0.44**
Micro-F1	85.32±0.27	86.48±0.33	82.48±0.66	83.29±0.41	91.54±0.29	49.62±0.51	92.82±0.42	90.26±0.34	**94.88±0.28**
Acc	79.38±0.93	81.10±0.61	80.74±0.62	82.23±0.74	90.61±0.34	74.33±0.18	92.14±0.61	91.23±0.14	**94.55±0.32**
Recall	95.64±0.15	93.85±0.37	89.04±0.37	88.56±0.58	97.26±0.11	49.68±0.30	95.83±0.22	94.92±0.30	**99.24±0.17**

However, they aim to learn the shared information on hypothesis spaces when addressing multiple tasks with different modalities. For the lifelong learning models, **ELLA** [1] first merges the features from different modalities into one high dimensional vector, and then explore a shared dictionary across different tasks. **lslMTMV** [43] constructs a decision function in the shared latent space across different modalities and tasks. **rLM^{2}L** [27] employs the robust multi-task learning formulation to explore the shared representations across different views. **L^{2}HMT** [31] constructs one common knowledge library for multiple modalities (*i.e.*, visual, tactile and auditory modalities) to store experience information of learned robotic tasks.

5.3. Performance comparison

In this subsection, as presented inTables 1–3, comparison experiments between our LVTL model and several state-of-the-art models are conducted to justify the superiority of our model. From Tables 1–3, we have the following conclusions: 1) Our LVTL model outperforms other competing methods especially for the lifelong learning approaches [1,27,31,43], since the modality-specific library and the sparse constraint based modality-invariant space could effectively explore the underlying intra-modality correlations and cross-modality complementary information. 2) Due to the lack of exploration about shared cross-modality representations, the multi-task learning methods [40,41] achieve worse performance. 3) Our model has more powerful continuous learning ability than multi-task multi-view approaches [42,44] by incorporating modality-specific library to preserve experience knowledge for each modality. 4) The large distribution divergence across visual and tactile modalities makes other competing methods difficult to capture shared discriminative representations across different modalities, while our LVTL model with a modality consistency regularizer could effectively alleviate it.

5.4. Qualitative analysis about computational costs

This subsection presents the experimental comparisons of optimization costs (*i.e.*, computational time) between our LVTL model and other competing comparison methods. As shown in Table 4, our model is computationally efficient to address a sequence of consecutive robotic tasks under the lifelong learning manner. Although multi-task learning models [40,41] consume less computational time, their prediction performances for lifelong learning robotic tasks are significantly worse than our model, as presented in Tables 1–3. More importantly, it validates that our proposed model with the efficient optimization procedure could be effectively applied into some real-world visual-tactile applications, when compared with other lifelong leaning approaches [1,27,31,43].

5.5. Effect of the number of learned tasks

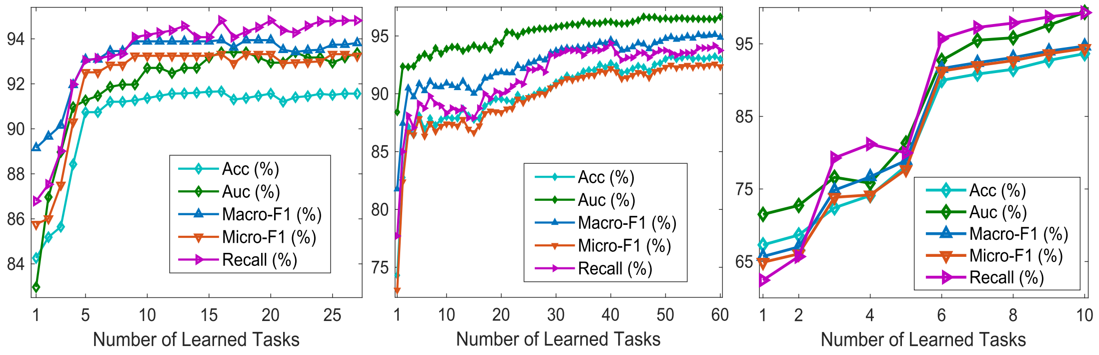
This subsection investigates whether the number of visual-tactile learning tasks affects the performance of our LVTL model when new robotic learning tasks arrive iteratively, as depicted in Fig. 2. From the presented results on all datasets, we can observe that the performances of our model in terms of five metrics are improving gradually as the number of learned tasks increasing continuously. More importantly, it validates that the proposed modality-specific knowledge library could effectively accumulate the experience knowledge for each modality to achieve lifelong learning. Furthermore, the evaluation curves in Fig. 2 strongly demonstrate the convergence of our LVTL model under the lifelong learning settings.

5.6. Effect of the regularizers P^t and Q^t

Figs. 3–4 report the effects of regularizers P^t and Q^t on GelFoldFabric, PHAC-2 and SVHN datasets by fixing γ_m as 0.001 and varying $\{\lambda_1, \lambda_2\}$ in range of $\{10^{-4}, 10^{-3}, 10^{-2}, 0.1, 1, 10\}$. From the presented results in Figs. 3–4, we can observe that our model achieves

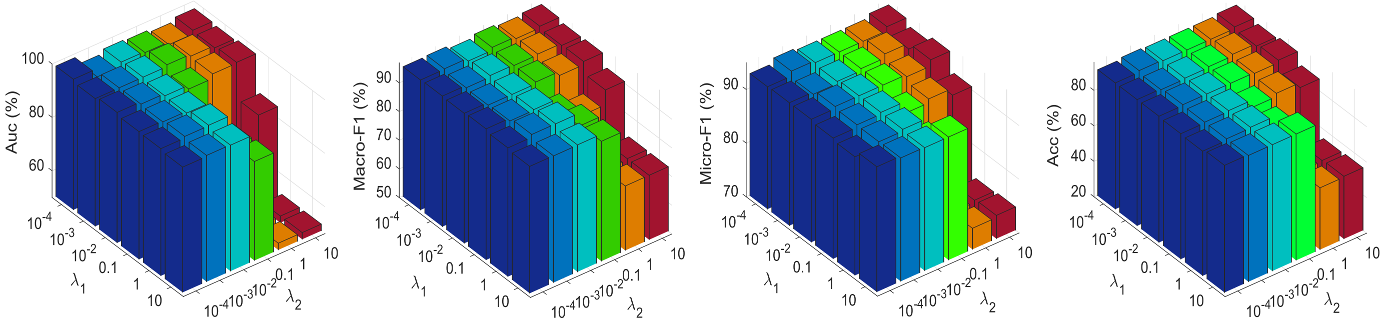
Table 4
Comparisons of computational time in terms of second (mean ± standard deviation) on benchmark datasets.

Dataset	MTFL [40]	rTGRT [41]	MAMUDA [42]	MFM [44]	ELLA [1]	lslMTMV [43]	rLM^{2}L [27]	L^{2}HMT [31]	Ours
PHAC-2	1.64±0.18	3.15±0.42	23.15±1.02	32.34±0.95	28.72±0.86	4.74±1.31	9.37±1.14	13.74±0.89	6.43±1.05
GelFoldFabric	4.30±0.62	7.41±0.35	46.57±1.16	89.43±1.24	53.47±1.83	16.79±0.62	22.74±0.83	32.72±0.68	19.56±0.71
SVHN	15.11±1.18	17.92±0.93	358.12±1.35	464.45±2.34	11.75±0.93	13.69±0.39	12.72±0.94	19.77±1.15	14.02±0.85
Average Time	7.02±0.66	9.49±0.57	142.61±1.18	195.41±1.51	31.31±1.21	11.74±0.77	14.94±0.97	22.08±0.91	13.34±0.87



(a) PHAC-2 (b) GelFoldFabric (c) SVHN

Fig. 2. Effect investigations about the number of learned visual-tactile tasks on PHAC-2 (a), GelFoldFabric (b) and SVHN (c) datasets.



(a) Auc (b) Macro-F1 (c) Micro-F1 (d) Acc

Fig. 3. Parameter investigations about $\{\lambda_1, \lambda_2\}$ on GelFoldFabric dataset in terms of Auc (a) and Macro-F1 (b), and PHAC-2 dataset in terms of Micro-F1 (c) and Acc (d).

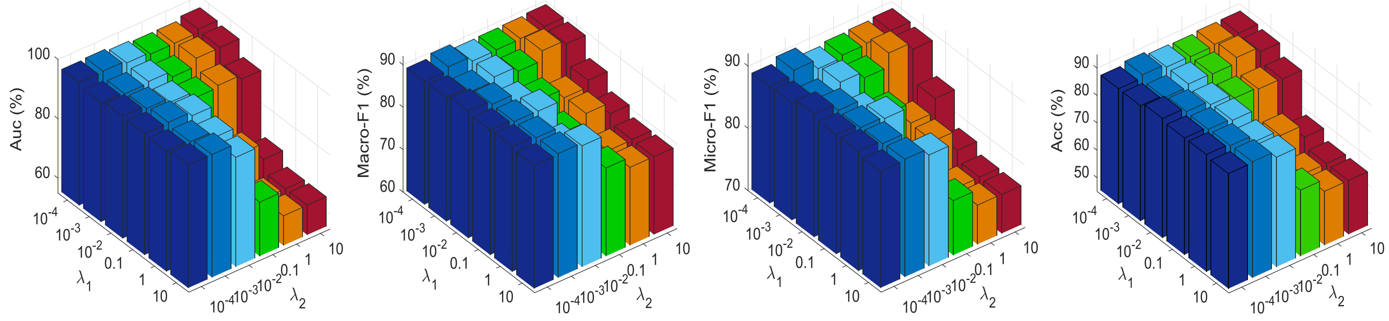
stable performance on several benchmark datasets in terms of all evaluation metrics, even though there is a large selection range of parameters $\{\lambda_1, \lambda_2\}$. When λ_1 and λ_2 are both large, the performance of our LVTL model decreases significantly since it pays more attention to highlight complementary information across modalities while neglecting the underlying modality-specific knowledge. Moreover, Figs. 3–4 also validate that our model could capture the complementary cross-modality correlations and identify the contributions of each modality for new robotic visual-tactile task.

5.7. Necessity of visual and tactile modalities

To verify the necessity of visual and tactile modalities in the lifelong robotic learning, we conduct some qualitative comparison experiments between our LVTL model and the lifelong learning method ELLA [1] that only accesses visual or tactile modality information. As shown in Fig. 5, the lifelong learning performance on PHAC-2 and GelFoldFabric datasets degrades significantly with respect to four metrics, when any one of modality information (visual or tactile modality) is unavailable. Obviously, it illustrates that both visual and tactile modalities play an essential role in compensating the misleading information of each modality for lifelong robotic learning. Our LVTL model could capture underlying cross-modality relationships and identify the contributions of each modality to improve the prediction performance for new visual-tactile task, by incorporating the sparse constraint based modality-invariant space. Furthermore, our proposed model could learn a sequence of robotic visual-tactile tasks continuously by exploring the intrinsic complementary knowledge while neglecting redundant information across visual and tactile modalities.

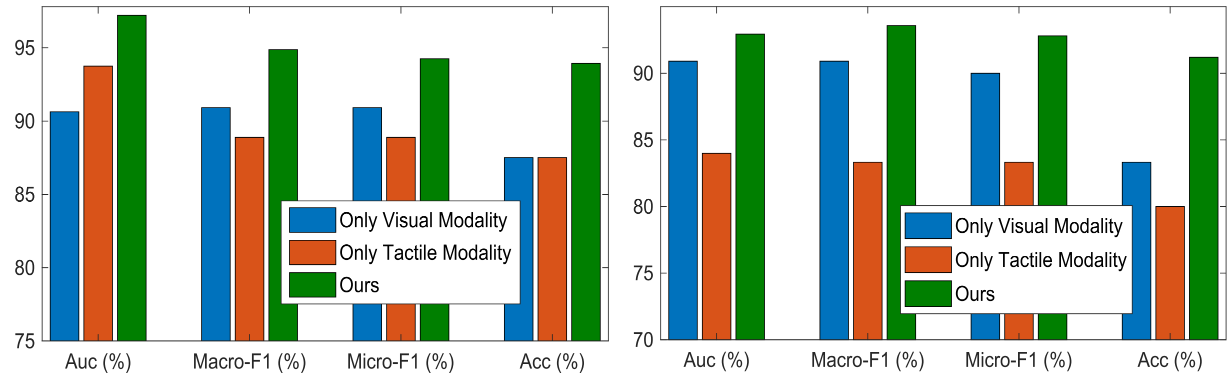
5.8. Convergence analysis

The optimization procedure of our proposed model consists of two components, *i.e.*, the optimization of variables $\{S^t, L_m\}$, as summarized in Algorithm 2. This subsection detailedly introduces the convergence analysis of our LVTL model with respect to the optimization of S^t (*i.e.*, Eq. (11)) by conducting experiments on PHAC-2, GelFoldFabri and SVHN datasets, since the optimization formulation for variable L_m has closed solution. As the convergence curves presented in Fig. 6, we can notice that the objective function of solving S^t (*i.e.*, Eq. 11) efficiently converges to a stable value across different robotic visual-tactile tasks. It validates that our model achieves stable convergence with respect to the variable S^t, and demonstrates the effectiveness of our optimization procedure presented in Algorithm 1 under the lifelong learning manner.



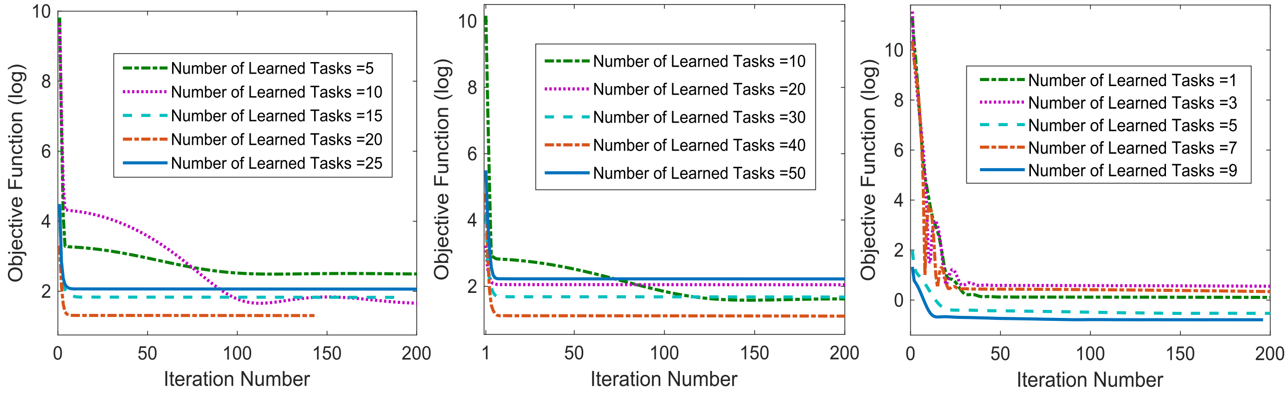
(a) Auc (b) Macro-F1 (c) Micro-F1 (d) Acc

Fig. 4. Parameter investigations about $\{\lambda_1, \lambda_2\}$ on SVHN dataset in terms of Auc (a), Macro-F1 (b), Micro-F1 (c) and Acc (d).



(a) PHAC-2 Dataset (b) GelFoldFabric Dataset

Fig. 5. Qualitative analysis about the necessity of visual and tactile modalities on PHAC-2 (a) and GelFoldFabri (b) datasets.



(a) PHAC-2 (b) GelFoldFabric (c) SVHN

Fig. 6. Convergence analysis of our LVTL model with respect to the variable S^t on PHAC-2 (a), GelFoldFabri (b) and SVHN (c) datasets.

5.9. Effect of the size of modality-specific library

In this subsection, PHAC-2 dataset is employed as an example to investigate the effect of the size k of modality-specific knowledge library, as depicted in Fig. 7. In the experiments, we tune k in the range of $[1, 27]$, while fixing the parameters $\{\lambda_1, \lambda_2, \gamma_m\}$. From the depicted curves in Fig. 7, we can notice that the performance of our proposed LVTL model is the best in terms of all evaluation metrics when k is selected as 2 or 3. Besides, the performance of our model decreases distinctly when the size k of modality-specific knowledge library increases gradually. The intuitive explanation for this phenomenon is that the modality-specific library with a large value of k would capture more redundant information and noisy representations across visual and tactile modalities. More importantly, Fig. 7 effectively illustrates that our model could highlight underlying complementary knowledge while neglecting irrelevant representations across modalities by incorporating the modality-specific library with appropriate size.

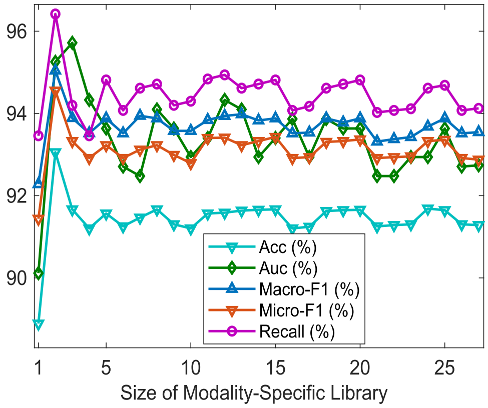
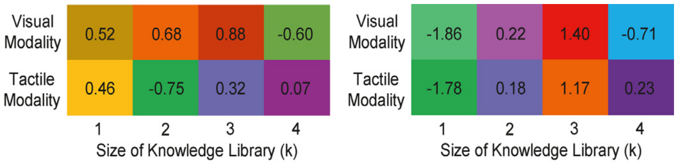


Fig. 7. Effect investigation about the size of modality-specific knowledge library on PHAC-2 dataset.

Table 5
Performance improvement (%) about the modality consistency regularizer in terms of five metrics (mean ± standard deviation) on benchmark datasets.

Dataset	Auc	Macro-F1	Micro-F1	Acc	Recall
PHAC-2	0.67±0.03	0.86±0.14	0.81±0.07	0.85±0.09	0.80±0.12
GelFoldFabric	0.82±0.11	0.59±0.07	0.69±0.10	0.91±0.17	0.72±0.08
SVHN	0.74±0.05	0.63±0.08	0.61±0.05	0.86±0.14	0.77±0.04



(a) PHAC-2: Values of S^{27} (b) GelFoldFabric: Values of S^{60}

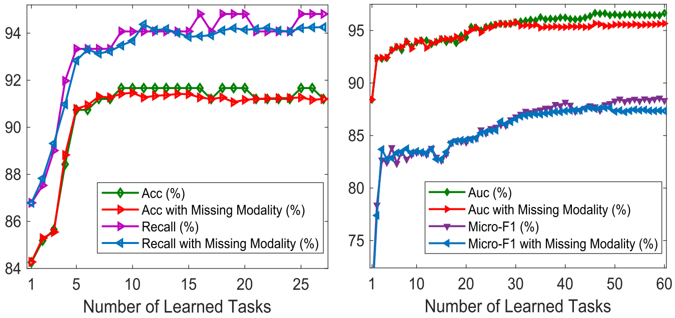
Fig. 8. Visualization of quantified contributions of visual and tactile modalities on PHAC-2 (S^{27}) and GelFoldFabri (S^{60}) datasets.

5.10. Effect of modality consistency regularizer

To investigate the effectiveness of the modality consistency regularizer, as shown in Table 5, we conduct ablation studies on benchmark datasets with respect to the modality consistency regularizer. Specifically, Table 5 presents the performance improvement of our LVTL model with the modality consistency regularizer (*i.e.*, $\gamma_m = 0.001$), when compared with the model without the modality consistency regularizer (*i.e.*, $\gamma_m = 0$). In Table 5, our proposed LVTL model has significant performance improvement in terms of all evaluation metrics when optimized with the modality consistency regularizer. Furthermore, Table 5 validates that it could ensure the semantic consistency between different modality-specific libraries while aligning the heterogeneous distribution gap across visual and tactile modalities. The modality consistency regularizer effectively encourages the samples from visual and tactile modalities to share the common knowledge bases for different modality-specific libraries.

5.11. Quantified contributions of visual and tactile modalities

To better understand the different contributions of visual and tactile modalities, as shown in Fig. 8, we visualize the values of $S^{27} \in \mathbb{R}^{k\times M}$ and $S^{60} \in \mathbb{R}^{k\times M}$ from the 27th task of PHAC-2 dataset and the 60th task of GelFoldFabric dataset, respectively. For the last learned visual-tactile perception task, S^{27} and S^{60} illustrate more correct view correlations, as sufficient semantic knowledge could be transferred into modality-specific libraries. From the presented results in Fig. 8, we realize that different modalities have different contributions for new visual-tactile perception task. More importantly, our proposed LVTL model could not only capture the complementary inter-modalities correlations, but also identify the task-specific modalities for each new task.



(a) PHAC-2 Dataset   (b) GelFoldFabric Dataset

Fig. 9. Effect investigations about missing visual modality or tactile modality on PHAC-2 (a) and GelFoldFabric (b) datasets.

5.12. Effect of missing modalities

In this subsection, we present evaluation experiments on PHAC-2 and GelFoldFabric datasets to investigate the performance of our LVTL model when any one of visual and tactile modalities is missing, as depicted in Fig. 9. As for the experimental configurations, one of visual and tactile modalities is randomly missing for the last 17 tasks in PHAC-2 dataset and the last 30 tasks in GelFoldFabric dataset, while the rest of tasks are with both visual and tactile modalities. From the depicted curves in Fig. 9, we realize that our proposed LVTL model achieves little performance degradation for the benchmark datasets with missing modalities. This observation validates that the modality-specific knowledge library could capture common intra-modality knowledge across different tasks to explore a modality-invariant space, which effectively tackles some challenging visual-tactile perception scenarios with missing modalities.

6. Conclusion

In this paper, we develop a new Lifelong Visual-Tactile Learning (LVTL) model to learn a sequence of robotic visual-tactile perception tasks continuously, which effectively explores the intrinsic intra-modality and cross-modality correlations. To be specific, we construct a modality-specific knowledge library for each modality to capture common intra-modality characterizations among different tasks, and utilize an auto-encoder mechanism to mitigate intra-modality mapping gap between semantic and feature spaces. Meanwhile, a sparse constraint based modality-invariant space is developed to highlight underlying cross-modality correlations and quantify the importance of each modality for new visual-tactile task. With a modality consistency regularizer to align the heterogeneous distributions across modalities, our LVTL model could ensure the semantic consistency between different modality-specific knowledge libraries. Evaluation experiments on several representative visual-tactile datasets strongly justify that our LVTL model performs significantly better than existing competing approaches with about 1.16%∼15.36% performance improvement for consecutive lifelong visual-tactile perception tasks. Our work is an earlier exploration about robotic visual-tactile cross-modality learning under the lifelong learning manner. A large number of robotic vision fields such as scene understanding, visual question answering, SLAM and underwater object detection would benefit from our work, due to the urgent requirement for robotic lifelong learning capacity. The performance of our LVTL model may suffer from the appearance shifts of robotic operation environments such as changes in weather and lighting conditions. In the future, we will improve the generalization performance of our model to perform well in various robotic visual-tactile perception systems with appearance shift and some new unknown classes (*i.e.*, open set domain adaptation [46–49]), when there is a large domain divergence between different application systems.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported in part by the National Key Research and Development Program of China under Grant 2019YFB1310300; in part by the National Nature Science Foundation of China under Grant 61821005, and 62003336; in part by National Postdoctoral Innovative Talents Support Program (BX20200353); and in part by the Nature Foundation of Liaoning Province of China under Grant 2020-KF-11-01.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.patcog.2021.108176.

References

[1] P. Ruvolo, E. Eaton, ELLA: an efficient lifelong learning algorithm, in: 30th International Conference on Machine Learning, ICML 2013, 2013.

[2] G. Sun, Y. Cong, X. Xu, Active lifelong learning with "watchdog", in: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2–7, 2018, 2018, pp. 4107–4114.

[3] S. Thrun, T.M. Mitchell, Lifelong robot learning, Rob. Auton. Syst. 15 (1) (1995) 25–46. The Biology and Technology of Intelligent Autonomous Agents.

[4] E. Einhorn, H. Gross, Generic 2D/3D SLAM with NDT maps for lifelong application, in: 2013 European Conference on Mobile Robots, 2013, pp. 240–247.

[5] A.J. Glover, G.F. Wyeth, Toward lifelong affordance learning using a distributed markov model, IEEE Trans. Cognit. Dev. Syst. 10 (1) (2018) 44–55.

[6] G. Sun, Y. Cong, J. Liu, L. Liu, X. Xu, H. Yu, Lifelong metric learning, IEEE Trans. Cybern. 49 (8) (2019) 3168–3179.

[7] G. Sun, Y. Cong, Y. Zhang, G. Zhao, Y. Fu, Continual multiview task learning via deep matrix factorization, IEEE Trans. Neural Netw. Learn. Syst. (2020) 1–12.

[8] J. Young, V. Basile, L. Kunze, E. Cabrio, N. Hawes, Towards lifelong object learning by integrating situated robot perception and semantic web mining, in: Proceedings of the Twenty-Second European Conference on Artificial Intelligence, in: ECAI'16, IOS Press, NLD, 2016, pp. 1458–1466.

[9] R. Xia, J. Jiang, H. He, Distantly supervised lifelong learning for large-scale social media sentiment analysis, IEEE Trans. Affect. Comput. 8 (4) (2017) 480–491.

[10] H. Liu, Y. Yu, F. Sun, J. Gu, Visual-tactile fusion for object recognition, IEEE Trans. Autom. Sci. Eng. 14 (2) (2017) 996–1008.

[11] W. Yuan, S. Wang, S. Dong, E. Adelson, Connecting look and feel: associating the visual and tactile properties of physical materials, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[12] D. Isele, M. Rostami, E. Eaton, Using task features for zero-shot knowledge transfer in lifelong learning, in: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, in: IJCAI'16, AAAI Press, 2016, pp. 1620–1626.

[13] A. Rannen, R. Aljundi, M.B. Blaschko, T. Tuytelaars, Encoder based lifelong learning, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017.

[14] H. Liu, F. Sun, Robotic Tactile Perception and Understanding: A Sparse Coding Method, Springer, 2018.

[16] J. Illonen, J. Bohg, V. Kyrki, 3-D object reconstruction of symmetric objects by fusing visual and tactile sensing, Int. J. Rob. Res. 33 (2) (2013) 321–341.

[17] S. Wang, J. Wu, X. Sun, W. Yuan, W.T. Freeman, J.B. Tenenbaum, E.H. Adelson, 3D shape perception from monocular vision, touch, and shape priors, in: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2018, Madrid, Spain, October 1–5, 2018, IEEE, 2018, pp. 1606–1613.

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

[19] Y. Li, J.-Y. Zhu, R. Tedrake, A. Torralba, Connecting touch and vision via cross-modal prediction, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[20] J. Lee, D. Bollegala, S. Luo, "Touching to See" and "Seeing to Feel": robotic cross-modal sensory data generation for visual-tactile perception, in: 2019 International Conference on Robotics and Automation (ICRA), 2019, pp. 4276–4282.

[21] S. Thrun, J. O'Sullivan, Discovering structure in multiple learning tasks: the TC algorithm, in: L. Saitta (Ed.), Proceedings of the 13th International Conference on Machine Learning ICML-96, Morgen Kaufmann, San Mateo, CA, 1996.

[22] A. Kumar, H. Daumé, Learning task grouping and overlap in multi-task learning, in: Proceedings of the 29th International Conference on International Conference on Machine Learning, in: ICML'12, Omnipress, Madison, WI, USA, 2012, pp. 1723–1730.

[23] A. Pentina, C.H. Lampert, Lifelong learning with non-i.i.d. tasks, in: C. Cortes, N.D. Lawrence, D.D. Lee, M. Sugiyama, R. Garnett (Eds.), Advances in Neural Information Processing Systems 28, Curran Associates, Inc., 2015, pp. 1540–1548.

[24] J. Cao, Z. Bu, Y. Wang, H. Yang, J. Jiang, H. Li, Detecting prosumer-community groups in smart grids from the multiagent perspective, IEEE Trans. Syst. Man Cybern. 49 (8) (2019) 1652–1664.

[25] Z. Bu, H. Li, J. Cao, Z. Wang, G. Gao, Dynamic cluster formation game for attributed graph clustering, IEEE Trans. Cybern. 49 (1) (2019) 328–341.

[26] C. Zhang, H. Fu, S. Liu, G. Liu, X. Cao, Low-rank tensor constrained multiview subspace clustering, in: 2015 IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1582–1590.

[27] G. Sun, Y. Cong, J. Li, Y. Fu, Robust lifelong multi-task multi-view representation learning, in: 2018 IEEE International Conference on Big Knowledge (ICBK), 2018, pp. 91–98.

[29] M. Hawasly, S. Ramamoorthy, Lifelong transfer learning with an option hierarchy, in: 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2013, pp. 1341–1346.

[30] W. Zheng, H. Liu, F. Sun, Lifelong visual-tactile cross-modal learning for robotic material perception, IEEE Trans. Neural Netw. Learn. Syst. (2020) 1–12.

[31] H. Liu, F. Sun, B. Fang, Lifelong learning for heterogeneous multi-modal tasks, in: 2019 International Conference on Robotics and Automation (ICRA), 2019, pp. 6158–6164.

[32] Y. Cong, J. Yuan, J. Liu, Abnormal event detection in crowded scenes using sparse representation, Pattern Recognit. 46 (7) (2013) 1851–1864.

[33] J. Bolte, S. Sabach, M. Teboulle, Proximal alternating linearized minimization for nonconvex and nonsmooth problems, Math. Program. 146 (2014) 459–494.

[34] A. Beck, M. Teboulle, A fast iterative shrinkage-thresholding algorithm with application to wavelet-based image deblurring, in: 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, 2009, pp. 693–696.

[35] J. Liu, J. Ye, Efficient L1/Lq norm regularization, 2010arXiv preprint arXiv:1009.4766

[36] I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Proceedings of the 27th International Conference on Neural Information Processing Systems, vol. 2, 2014, pp. 2672–2680.

[37] Y. Gao, L.A. Hendricks, K.J. Kuchenbecker, T. Darrell, Deep learning for tactile understanding from visual and haptic data, in: Proceedings of the IEEE International Conference on Robotics and Automation, 2016, pp. 536–543.

[38] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: International Conference on Learning Representations, 2015.

[39] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, A.Y. Ng, Reading digits in natural images with unsupervised feature learning, NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011, 2011.

[40] P. Gong, J. Ye, C. Zhang, Robust multi-task feature learning, in: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, in: KDD '12, Association for Computing Machinery, New York, NY, USA, 2012, pp. 895–903.

[41] Y. Yao, J. Cao, H. Chen, Robust task grouping with representative tasks for clustered multi-task learning, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery Data Mining, in: KDD '19, Association for Computing Machinery, New York, NY, USA, 2019, pp. 1408–1417.

[42] X. Jin, F. Zhuang, H. Xiong, C. Du, P. Luo, Q. He, Multi-task multi-view learning for heterogeneous tasks, in: Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, in: CIKM '14, Association for Computing Machinery, New York, NY, USA, 2014, pp. 441–450.

[43] X. Li, S.N. Chandrasekaran, J. Huan, Lifelong multi-task multi-view learning using latent spaces, in: 2017 IEEE International Conference on Big Data (Big Data), 2017, pp. 37–46.

[44] C.-T. Lu, L. He, W. Shao, B. Cao, P.S. Yu, Multilinear factorization machines for multi-task multi-view learning, in: Proceedings of the Tenth ACM International Conference on Web Search and Data Mining, in: WSDM '17, Association for Computing Machinery, New York, NY, USA, 2017, pp. 701–709.

[45] T. Zhang, Y. Cong, G. Sun, J. Dong, Visual-tactile fused graph learning for object clustering, IEEE Trans. Cybern. (2021).

[46] Y. Zhang, F. Liu, B. Yuan, G. Zhang, J. Lu, Clarinet: A One-step Approach Towards Budget-friendly Unsupervised Domain Adaptation, in: Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, ijcai.org, 2021.

[47] L. Zhong, Z. Fang, F. Liu, J. Lu, et al., How does the Combined Risk Affect the Performance of Unsupervised Domain Adaptation Approaches? AAAI, AAAI Press, 2021.
[48] Z. Fang, J. Lu, F. Liu, J. Xuan, G. Zhang, Open Set Domain Adaptation: Theoretical Bound and Algorithm, IEEE Transactions on Neural Networks and Learning Systems (2020).
[49] Y. Zhang, F. Liu, Z. Fang, G. Zhang, J. Lu, Learning from a Complementary-label Source Domain: Theory and Algorithms, IEEE Transactions on Neural Networks and Learning Systems (2020).
[50] Q. Wang, Z. Ding, Z. Tao, Q. Gao, Y. Fu, Generative Partial Multi-View Clustering With Adaptive Fusion and Cycle Consistency, IEEE Transactions on Image Processing 30 (2021) 1771–1783.
[51] J. Li, Z. Tao, Y. Wu, B. Zhong, Y. Fu, Large-Scale Subspace Clustering by Independent Distributed and Parallel Coding, IEEE Trans. Cybern. (2021).
[52] J. Dong, Y. Cong, G. Sun, B. Zhong, X. Xu, What Can Be Transferred: Unsupervised Domain Adaptation for Endoscopic Lesions Segmentation, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Publisher, 2020, pp. 4022–4031.

Jiahua Dong is currently a Ph. D candidate in State Key Laboratory of Robotics, Shenyang Institute of Automation, University of Chinese Academy of Sciences. He received the B.S. degree from Jilin University in 2017. His current research interests include computer vision, machine learning, transfer learning, domain adaptation and medical image processing.

Yang Cong (S'09-M'11-SM'15) is a full professor of Shenyang Institute of Automation, Chinese Academy of Sciences. He received the he B.Sc. de. degree from Northeast University in 2004, and the Ph.D. degree from State Key Laboratory of Robotics, Chinese Academy of Sciences in 2009. He was a Research Fellow of National University of Singapore (NUS) and Nanyang Technological University (NTU) from 2009 to 2011, respectively; and a visiting scholar of University of Rochester. His current research interests include image processing, compute vision, machine learning, multimedia, medical imaging, data mining and robot navigation.

Gan Sun (S'19) is an Assistant Professor in State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences. He received the B.S. degree from Shandong Agricultural University in 2013, the Ph.D. degree from State Key Laboratory of Robotics, Chinese Academy of Sciences in 2020. His current research interests include lifelong machine learning, multi-task learning, medical data analysis, deep learning and 3D computer vision.

Tao Zhang is currently working toward the Ph.D. degree in pattern recognition and intelligent systems at the State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang, China. His research interests include pattern recognition, image processing, tactile sensing and robotics.