Generative Partial Visual-Tactile Fused Object Clustering

Tao Zhang[1,2,3], Yang Cong[1]*, Gan Sun[1], Jiahua Dong[1,2,3], Yuyang Liu[1,2,3], Zhenming Ding[4]

[1]State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences
[2]Institutes for Robotics and Intelligent Manufacturing Chinese Academy of Sciences
[3]University of Chinese Academy of Sciences
[4]Department of Computer Science, Tulane University
zhangtaosia@gmail.com, yangcong81@gmail.com, sungan1412@gmail.com, dongjiahua1995@gmail.com,
liuyuyang@sia.cn, zding1@tulane.edu

Abstract

Visual-tactile fused sensing for object clustering has achieved significant progresses recently, since the involvement of tactile modality can effectively improve clustering performance. However, the missing data (*i.e., partial data*) issues always happen due to occlusion and noises during the data collecting process. This issue is not well solved by most existing partial multi-view clustering methods for the heterogeneous modality challenge. Naively employing these methods would inevitably induce a negative effect and further hurt the performance. To solve the mentioned challenges, we propose a Generative Partial Visual-Tactile Fused (*i.e.,* GPVTF) framework for object clustering. More specifically, we first do partial visual and tactile features extraction from the partial visual and tactile data, respectively, and encode the extracted features in modality-specific feature subspaces. A conditional cross-modal clustering generative adversarial network is then developed to synthesize one modality conditioning on the other modality, which can compensate missing samples and align the visual and tactile modalities naturally by adversarial learning. To the end, two pseudo-label based KL-divergence losses are employed to update the corresponding modality-specific encoders. Extensive comparative experiments on three public visual-tactile datasets prove the effectiveness of our method.

Introduction

Benefitting from the great progresses in visual-tactile fused sensing (Liu and Sun 2018; Luo et al. 2018; Lee, Bollegala, and Luo 2019), researchers (Zhang et al. 2020) attempt to focus on visual-tactile fused clustering (VTFC), which aims to group similar objects together in an unsupervised manner. An interesting example is that when robots employ visual and tactile information to explore unknown environment
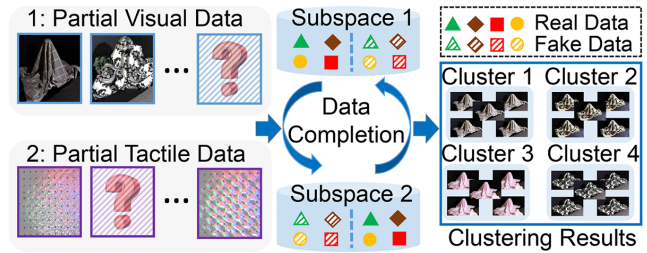
Figure 1: Diagram of our proposed method, which first encodes original partial visual and tactile data in modality-specific subspaces. Then, we do visual-tactile fused clustering after completing the missing data.

(e.g., many objects cluttered in an unstructured scene), recognizing the objects in this scene by collecting and annotating a lot of samples is time-consuming and expensive (Zhao, Wang, and Huang 2021; Wei et al. 2019; Zhao et al. 2020; Wei, Deng, and Yang 2020; Sun et al. 2020b). An alternative solution is to use unsupervised manner to group these objects. In this setting, previous VTFC methods provide a feasible solution by employing fused visual-tactile information in an unsupervised manner to group the objects with same identity into a same group (i.e., object clustering). Fusion visual-tactile information could improve the clustering performance effectively, since they can provide complementary information. Generally, most existing VTFC methods mainly utilize the idea of multi-view clustering (Dang et al. 2020; Hu, Shi, and Ye 2020; Hu, Yan, and Ye 2020), *e.g.*, Zhang et al. (Zhang et al. 2020) propose a VTFC model based on non-negative matrix factorization (NMF) as well as consensus clustering and achieve great progresses. As far as we know, this is the first work about visual-tactile fused clustering.

However, the task of VTFC has not been well settled due to the following challenges *i.e.*, **partial data** and **heterogeneous modality**. **Partial data:** Existing visual-tactile fused object clustering methods (Zhang et al. 2020) make a strong assumption that all the visual-tactile modalities well aligned and complete. However, visual-tactile data usually tend to

be incomplete in real world applications. For instance, when a robot grasps an apple, the visual information of the apple becomes unobservable due to the occlusion of a robot hand. Moreover, noises, signal loss and malfunction in the data collecting process might make the instance missing. For instance, in special situations (*e.g.,* underwater scenes), the visual can be easily missing due to turbidity of the water. These cases mentioned above lead to the incompleteness of multi-modality data, which further hurt the clustering performance. **Heterogeneous modality:** Most previous partial multi-view clustering methods use different feature description methods (*e.g.*, SIFT, LBP, HOG) to extract different view features for visual data, which are essentially homogeneous data. Therefore, directly employing these methods on heterogeneous data (*i.e.*, visual and tactile data) could induce a negative effect and even unsuccessful clustering task, since they ignore the distinct properties between visual and tactile modalities.

To solve these problems mentioned above, as shown in Figure 1, we propose a Generative Partial Visual-Tactile Fused (*i.e.*, GPVTF) framework for object clustering, which aims to obtain better clustering results by adopting generative adversarial learning as well as simple yet effective KL-divergence losses. Specifically, we first extract partial visual and tactile features from the raw input data, and employ two modality-specific encoders to project the extracted features into visual subspace and tactile subspace, respectively. Then visual (or tactile) conditional cross-modal clustering generative networks are trained to reproduce tactile (or visual) latent representations in the modality-specific subspaces. In this way, the our proposed approach is able to effectively leverage the complementary information, and learns the latent subspace level pairwise cross-modal knowledge among visual-tactile data. The conditional clustering generative adversarial networks can not only complete the missing data, but also force the heterogeneous modalities to be similar and further align them. With the well completed and aligned visual and tactile subspaces, we can obtain expressive representations of the raw visual-tactile data. Moreover, two pseudo label based fusion KL-divergence losses are employed to update the encoders, and further help obtaining better representations for better clustering performance. Finally, extensive experimental results on three real-world visual-tactile datasets prove the superiority of our proposed framework. We summarize the contributions of our work as follows:

- We put forward a Generative Partial Visual-Tactile Fused (GPVTF) framework for partial visual-tactile clustering. To our best knowledge, this is an earlier work about visual-tactile fused clustering, which tackles the problem of incomplete data.

- A conditional cross-modal clustering generative adversarial learning schema is encapsulated in our model to complete the missing data and align visual-tactile data, which can further help explore the shared complementary information among multi-modality data.

- We conduct comparisons and experiments with three benchmark real-world visual-tactile datasets, which show the superiority of the proposed GPVTF framework.

Related Work

Visual-Tactile Fused Sensing

Significant progresses have been made on visual-tactile fused sensing (Liu and Sun 2018) in recent years, *e.g.*, object recognition, cross-modal matching and object clustering. For example, Liu et al. (Liu et al. 2016) develop an effective fusion strategy for weakly paired visual-tactile data based on joint sparse coding, which makes great success in household object recognition. Wang et al. (Wang et al. 2018b) predict the shape prior of an object from a single color image and then achieve accurate 3D object shape perception by actively touching the object. Yuan et al. (Yuan et al. 2017) show that there is an intrinsic connection between visual and tactile modalities through the physical properties of materials. Li et al. (Li et al. 2019) uses a conditional generative adversarial network to generate pseudo visual (or tactile) outputs based on tactile (or visual) inputs, then expanding the generated data to classification tasks. Zhang et al. (Zhang et al. 2020) first propose a visual-tactile fusion object clustering framework base on non-negative matrix factorization (NMF). However, all of the methods assume that data are well aligned and complete, which is unrealistic in practical applications. Thus, we design the GPVTF framework to address these problems for object clustering in this paper.

Partial Multi-View Clustering

Partial multi-view clustering (Sun et al. 2020a; Li, Jiang, and Zhou 2014; Wang et al. 2020, 2018a), which provides a framework to solve the issue of incomplete (partial) input data, can be divided into two categories. The first category is based on traditional technique, such as NMF and kernel learning. For example, Li et al. (Li, Jiang, and Zhou 2014) propose a incomplete multi-view clustering framework by establishing a latent subspace based on NMF, where incomplete multi-view information is maximized. Shao et al. (Shao, Shi, and Philip 2013) propose a collective kernel learning method to complete missing data and then do clustering tasks. The second category utilizes generative adversarial networks (GANs) to complete the missing data, for the reason that GANs can align heterogeneous data and complete partial data (Dong et al. 2020, 2019; Yang et al. 2020; Jiang et al. 2019). For instance, Xu et al. (Xu et al. 2019) propose an adversarial incomplete multi-view clustering method, which performs missing data inference via GANs and learns the common latent subspace of multi-view data simultaneously. All the methods mentioned above are developed for homogeneous data, they ignore the huge gap between heterogeneous data (*i.e.*, visual and tactile data).

The Proposed Method

In this section, the proposed Generative Partial Visual-Tactile Fused (GPVTF) framework is presented in detail, together with its implementation.

Details of the Model Pipeline

Given the visual-tactile data V and T, where V denotes the visual data (*i.e.*, RGB images) and T denotes the tactile data.
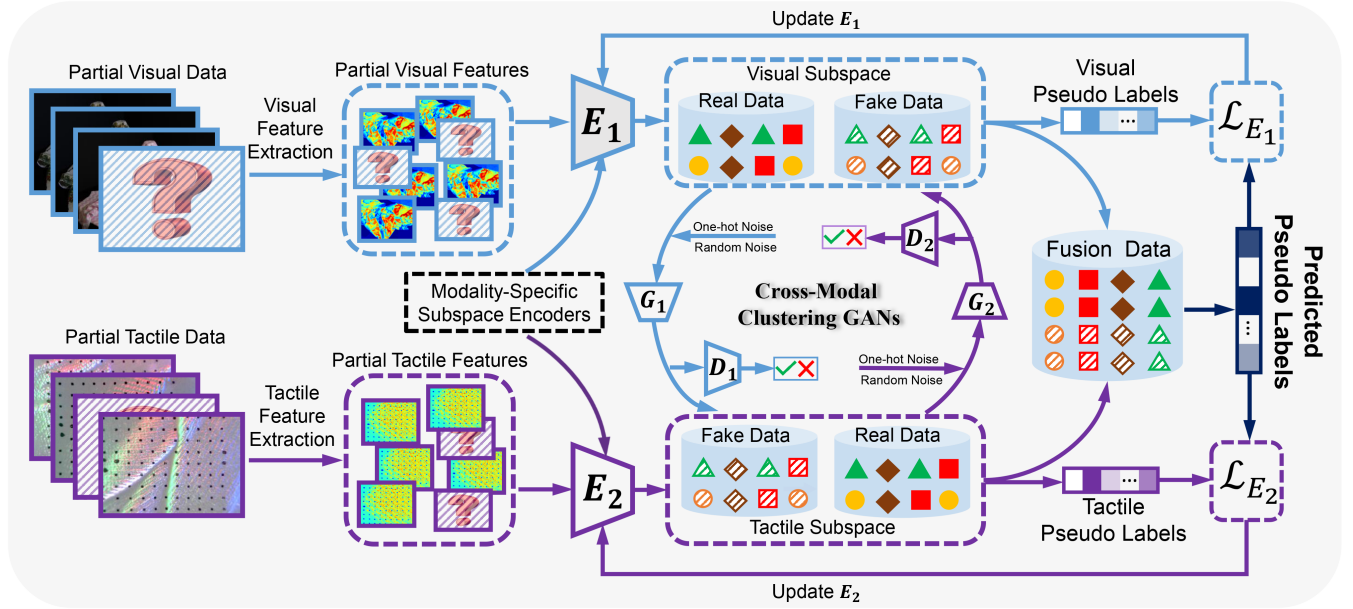
Figure 2: Illustration of the proposed generative partial visual-tactile fused object clustering framework. Firstly, partial visual and tactile features are extracted from the raw partial visual and tactile data. Then two modality-specific encoders, *i.e.*, visual encoder $E_1(\cdot)$ and tactile encoder $E_2(\cdot)$ are introduced to obtain distinctive representations in visual subspace and tactile subspace. Two cross-modal clustering generators $G_1(\cdot)$ and $G_2(\cdot)$ generate representations conditionally based on the other subspace, which could not only pull the distance between the visual and tactile subspaces but also complete the missing items. Finally, both real and generated fake representations are fused to predict the clustering labels. Meanwhile, the modality-specific encoders are updated by the KL-divergence losses $\mathcal{L}_{E_1}$ and $\mathcal{L}_{E_2}$, which are calculated by the predicted pseudo labels.

Noticing that the visual and tactile data collected from different tactile sensors lie in different data spaces. Our proposed GPVTF model consists of two partial feature extraction processes, *i.e.,* visual feature extraction and tactile feature extraction, which learn partial visual features $X_n^{(1)} \in \mathbb{R}^{d_1 \times n}$ from V and tactile features $X_n^{(2)} \in \mathbb{R}^{d_2 \times n}$ from T, where d_1 and d_2 are the feature dimensions and n is the number of samples; two modality-specific encoders, $E_1(\cdot)$ and $E_2(\cdot)$; two generators, $G_1(\cdot)$ and $G_2(\cdot)$ and their corresponding discriminators, $D_1(\cdot)$ and $D_2(\cdot)$; two KL-divergence based losses, as illustrated in Figure 2. More details are provided in the following sections. Particularly, since each dataset has different feature extraction processes, the details of these processes are given in the "Experiments" section.

Encoders and Clustering Module: Modality-specific encoders $E_1(\cdot)$ and $E_2(\cdot)$ are introduced to project both partial visual and tactile features into the modality-specific subspaces, *i.e.,* visual subspace and tactile subspace, respectively. Specifically, in the modality-specific subspaces, the learn latent subspace representations are learned via $Z_n^{(m)} = E_m(X_n^{(m)}; \theta_{E_m})$, where $m = 1$ denotes the visual modality, $m = 2$ denotes the tactile modality, and θ_{E_m} denote the network parameters of the m-th encoder. Then the fused representations (*i.e.*, $m = 3$) can be gained by:

$$Z_n^{(3)} = (1 - \alpha)Z_n^{(1)} + \alpha Z_n^{(2)}, \quad (1)$$

where $\alpha > 0$ is the weighting coefficient that balances the ratio of tactile and visual modalities. Next, the K-means

method is employed on $Z_n^{(m)}$ to get the initial clustering centers $\{\mu_j^{(m)}\}_{j=1}^k$, where k is the number of clusters[1]. Inspired by (Xie, Girshick, and Farhadi 2016), we employ Student's t-distribution to measure the similarity of latent subspace representations $Z_n^{(m)}$ and the clustering center $\mu_j^{(m)}$:

$$q_{nj}^{(m)} = \frac{(1 + \|Z_n^{(m)} - \mu_j^{(m)}\|^2/\gamma)^{-\frac{2}{\gamma+1}}}{\sum_{j'}(1 + \|Z_n^{(m)} - \mu_{j'}^{(m)}\|^2/\gamma)^{-\frac{\gamma+1}{2}}}, \quad (2)$$

where γ is the degrees of freedom of the Student's t-distribution and set to be 1 in this paper; $q_{nj}^{(m)}$ are the pseudo-labels, which denote the probability of assigning sample n to cluster j for the m-th modality.

To improve cluster compactness, we pay more attention to data points of which are assigned with high confidence, by obtaining the target distribution $p_{nj}^{(m)}$ as follows:

$$p_{nj}^{(m)} = \frac{q_{nj}^{(m)2}/\sum_n q_{nj}^{(m)}}{\sum_{j'} q_{nj'}^{(m)2}/\sum_n q_{nj}^{(m)}}. \quad (3)$$

Then the encoders are trained with fused KL-divergence

---

[1]Since we do clustering according to object identity, the k is set to be equal with the number of types of objects in the datasets. Specifically, k is set to be 53, 119 and 108 for **PHAC-2**, **GelFabric**, and **LMT** datasets, respectively.

losses, which are defined as follows:

$$\mathcal{L}_{E_m} = KL\big(P^{(m)}||Q^{(m)}\big) + \beta KL\big(P^{(3)}||Q^{(3)}\big)$$

$$= \sum_n \sum_j p_{nj}^{(m)} \log \frac{p_{nj}^{(m)}}{q_{nj}^{(m)}} + \beta \sum_n \sum_j p_{nj}^{(3)} \log \frac{p_{nj}^{(3)}}{q_{nj}^{(3)}}, \quad (4)$$

where $m = 1$ and $m = 2$ correspond to the losses of encoders $E_1(\cdot)$ and $E_2(\cdot)$, and β is a trade-off parameter. The encoders are implemented by a two-layer fully-connected network.

Conditional Cross-Modal Clustering GANs: Noticing that the gap between visual and tactile modalities is very large since their frequency, format and receptive field are quite different. Thus, directly employing GANs in the original space $X_n^{(m)}$ might increase the difficulty of training or even lead to non-convergence. To address this challenge, we develop a conditional cross-modal clustering GANs, which generates one latent space conditional on the other latent space. Specifically, the conditional cross-modal cluster GANs including $G_m(\cdot)$ and $D_m(\cdot)$, where $G_m(\cdot)$ competes with $D_m(\cdot)$ to generate samples as real as possible, and the loss function is given as:

$$\mathcal{L}_{G_{md}} = -E_{\omega \sim P_\omega(\omega)} \log\big(1 - D_m(G_m(\omega|Z_n^{(m)}))\big), \quad (5)$$

where ω is the noise matrix. Noticing that our goal is clustering rather than generation, a prior that consists of normal random variables cascaded with one-hot noise is sampled, which is different from tradition GANs. More specifically, $\omega = (\omega_n, \omega_c)$, $\omega_n \sim N(0, \sigma^2 I_{dn})$, $\omega_c = e_k$, e_k is the k-th elementary vector in $\mathbb{R}^k$ and k is the number of clusters. We choose $\sigma = 0.1$ in all our experiments. By this way, a non-smooth geometry latent subspace is created, and $G_m(\cdot)$ can generate more distinctive and robust representations which are beneficial for clustering performance, *i.e.*, not only the gap between visual and tactile modalities can be mitigated but also the missing data are completed naturally.

Moreover, since training the GANs in Eq. (5) is not trivial (Wang et al. 2019), a regularizer, which forces the real samples and the generated fake samples to be similar, is introduced to obtain stable generative results, which can be defined as:

$$\mathcal{L}_{G_{ms}} = E_{\omega \sim P_\omega(\omega)}\big(\|G_m(\omega|Z_n^{(m)})) - Z_n^{(m)}\|^2\big). \quad (6)$$

Then, the overall loss function of $G_m(\cdot)$ is given as follows:

$$\mathcal{L}_{G_m} = \mathcal{L}_{G_{md}} + \lambda \mathcal{L}_{G_{ms}}, \quad (7)$$

where λ is a trade-off parameter which balances the two losses and is set to be 0.1 in this paper. $G_m(\cdot)$ is a three-layer network.

The discriminator $D_m(\cdot)$ is designed to discriminate the fake representations generated by $G_m(\cdot)$ and the real representations in the modality-specific subspaces. The object function for $D_m(\cdot)$ can be given as:

$$\mathcal{L}_{D_m} = E_{Z \sim P_Z(Z)} \log D_m(E_m(X_n^{(m)}; \theta_{E_m})) +$$
$$E_{\omega \sim P_\omega(\omega)} \log\big(1 - D_m(G_m(\omega|E_m(X_n^{(m)}; \theta_{E_m})))\big). \quad (8)$$

The proposed $D_m(\cdot)$ is mainly made up of a fully-connected layer with ReLU activation, a mini-batch layer (Salimans et al. 2016) that can increase the diversity of fake representations, a sigmoid function which outputs the fake-real possibility of input representations. Then, both the generated fake and real representations are fused. Thus, the fusion representations Eq. (1) can be modified to:

$$Z_n^{(3)} = (1 - \alpha)Z_n^{(1)} + \alpha Z_n^{(2)} + \sum_{m=1}^{2} \varphi_m Z_{\text{fake}}^{(m)}, \quad (9)$$

where φ_m is the weighting coefficients of real and the generated fake representations for the m-th modality, *i.e.*, $m = 1$ represents visual modality and $m = 2$ represents tactile modalities, respectively. The overall loss function of our model is summarized as follows:

$$\mathcal{L}_{total} = \min_{E_m, G_m} \max_{D_m} \mathcal{L}_{E_m} + \mathcal{L}_{G_m} + \mathcal{L}_{D_m}, \quad (10)$$

where $\mathcal{L}_{E_m}$ are the KL-divergence losses, $\mathcal{L}_{G_m}$ and $\mathcal{L}_{D_m}$ are the conditional cross-modal clustering GANs losses.

Algorithm 1 Training Process of the Proposed Framework

1: **Input:** Visual-tactile data: $\{V, T\}$. Number of clusters: k. The maximum number of iterations: MaxIter; hyper-parameters α, β, φ_1 and φ_2.
2: **Initialization:** Project $\{V, T\}$ into feature subspaces $\{X_n^{(1)}, X_n^{(2)}\}$. Initialize the parameters of the networks with Xavier initializer. Calculate the initial fusion representations $Z_n^{(3)}$ and the clustering centers $\{\mu_j^{(m)}\}_{j=1}^k$.
3: **for** iter $\leq$ MaxIter **do**
4:     Train the encoders $E_m(\cdot)$ with corresponding KL-divergence losses $\mathcal{L}_{E_m}, \forall m = 1, 2$.
5:     Train the generators $G_m(\cdot)$ with $\mathcal{L}_{G_m}, \forall m = 1, 2$.
6:     Train the discriminators $D_m(\cdot)$ with $\mathcal{L}_{D_m}, \forall m = 1, 2$.
7:     Update the fused representation $Z_n^{(3)}$ and clustering centers $\{\mu_j^{(m)}\}_{j=1}^k, \forall m = 1, 2$.
8: **end for**
9: Gain the updated fusion representation $Z_n^{(3)}$, fusion clustering centers $\{\mu_j^{(3)}\}_{j=1}^k$ and pseudo-labels $q_{nj}^{(3)}$.
10: Predict the clustering labels according to $q_{nj}^{(3)}$.
11: **return** Predicted cluster labels.

Training

The whole process of the proposed GPVTF framework is summarized as below.

Step 1 Initialization: We feed the partial visual and tactile features $X_n^{(1)}$ and $X_n^{(2)}$ into $E_1(\cdot)$ and $E_2(\cdot)$ to obtain the initial latent subspace representations $Z_n^{(m)}$. Then standard K-means method is applied on $Z_n^{(m)}$ to get the initial clustering centers $\{\mu_j^{(m)}\}_{j=1}^k, \forall m = 1, 2, 3$.

Step 2 Training encoders: Eq. (2) is employed to calculate the pseudo-labels $q_{nj}^{(m)}; p_{nj}^{(m)}$ and KL-divergence losses L_{E_m} are computed by Eq. (3) and Eq. (4), respectively. Then L_{E_m}

are fed to its corresponding Adam optimizers to train the encoders and the learning rates are set to be 0.0001.

Step 3 *Training conditional cross-modal clustering GANs*: In this step, we employ the generator losses, *i.e.*, Eq. (5) and Eq. (6) with Adam optimizers to update the parameters of the two generators and the learning rates are set to be 0.000003 and 0.000004 for $G_1(\cdot)$ and $G_2(\cdot)$, respectively. Next, the two discriminators $D_1(\cdot)$ and $D_2(\cdot)$ are optimized by Eq. (8) with Adam optimizers and the leaning rates are set to be 0.000001 both for $D_1(\cdot)$ and $D_2(\cdot)$. We update the generators five times while updating the discriminators once.

Step 4 After the framework is optimized, we feed original data to the model and then obtain the completed fusion representations $Z_n^{(3)}$ as well as the updated clustering centers $\{\mu_j^{(m)}\}_{j=1}^k$. Then the predicted clustering labels $q_{nj}^{(3)}$ are calculated by Eq. (2). Finally, we choose the maximum value of $q_{nj}^{(3)}$ as the predicted clustering labels. We implement the model with Tensorflow 1.12.0, and set the batch size to be 64. We summarize the overall training process of the proposed framework in **Algorithm 1**.

Experiments

In this section, the used datasets, comparison methods, evaluation metrics and experimental results are given.

Datasets and Partial Data Generation

PHAC-2 (Gao et al. 2016) dataset consists of color images and tactile signals of 53 household objects, where each object has 8 color images and 10 tactile signals. We use all the images and the first 8 tactile signals to build the initial paired visual-tactile dataset in this paper. The feature extraction process of the tactile modality is similar with (Gao et al. 2016; Zhang et al. 2020), and the visual features are extracted by AlexNet (Krizhevsky, Sutskever, and Hinton 2012), which is pre-trained on the ImageNet. After feature extraction, 4096-D visual and 2048-D tactile features are obtained. **LMT** (Zheng et al. 2016; Strese, Schuwerk, and Steinbach 2015) dataset consists of 10 color images and 30 haptic acceleration data of 108 different surface materials. The first 10 haptic acceleration data and all the images are used. We extract 1024-D tactile features similarly with (Liu, Sun, and Fang 2019) and 4096-D visual features by the pre-trained AlexNet. **GelFabric** (Yuan et al. 2017) dataset includes visual data (*i.e.,* color and depth images) and tactile data of 119 kind of different fabrics. Each fabric has 10 color images and 10 tactile images, which are used in this paper. Since both the visual and tactile data are image formats, we extract 4096-D visual and tactile features with pre-trained AlexNet. Some examples of the used datasets are given in Figure 3.

Partial data generation: The partial visual-tactile datasets are generated in a similar way with partial multi-view clustering settings, *e.g.*, Xu et al (Xu et al. 2019). Supposing that the number of all the visual and tactile samples is N in each dataset, we randomly select $\tilde{N}$ samples as the missing data points. Then, the Missing Rate (*i.e.*, $\mathcal{MR}$) can be defined as $\mathcal{MR} = \frac{\tilde{N}}{N}$.

PHAC-2 : (a) Visual Modality   PHAC-2 : (b) Tactile Modality

LMT : (a) Visual Modality   LMT : (b) Tactile Modality

GelFabric : (a) Visual Modality   GelFabric : (b) Tactile Modality
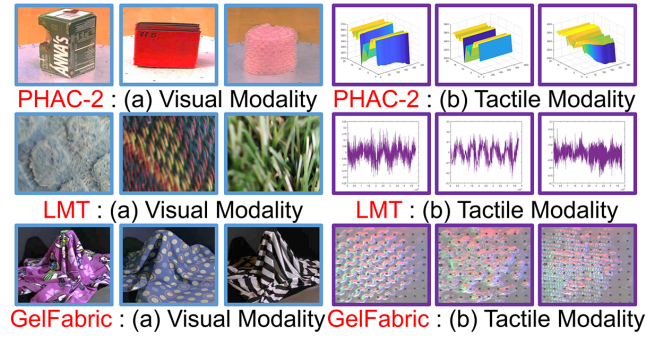
Figure 3: Example visual images and tactile data of the used datasets, *i.e.,* GelFabric dataset (Yuan et al. 2017), LMT dataset (Strese, Schuwerk, and Steinbach 2015; Strese et al. 2014), and PHAC-2 dataset (Gao et al. 2016). Intuitively, there are intrinsic differences between visual and tactile data. It is worth noting that the tactile signals in the PHAC-2 dataset consist of multiple components. We only visualize the electrode impedance component for simplicity. More details of these datasets can be found in their corresponding references.

Comparsion Methods and Evaluation Metrics

We compare our GPVTF model with the following baseline methods. We first employ standard spectral clustering methods on the modality-specific features, *i.e.*, visual features $X_n^{(1)}$ and tactile features $X_n^{(2)}$, which are termed as **SC1** and **SC2**. **ConcatPCA** concatenates feature vectors of different modalities via PCA and then performs standard spectral clustering. **GLMSC** (Zhang et al. 2018) proposes a subspace multi-view clustering model under the assumption that each single feature view originates from one comprehensive latent representations. **VTFC** (Zhang et al. 2020) is a pioneering work to incorporate visual modality with tactile modality in the object clustering tasks based on autoencoders and NMF. **IMG** (Zhao, Liu, and Fu 2016) does the incomplete multi-view clustering by transforming the original partial data to complete representations. **GRMF** (Wen et al. 2018) exploits the complementary and local information among all views and samples based on graph regularized matrix factorization. **UEAF** (Wen et al. 2019) performs missing data inference with locality-preserved constraint.

Evaluation Metrics: Two widely used clustering evaluation metrics, *i.e.*, Accuracy (ACC) and Normalized Mutual Information (NMI) are employed to assess the effectiveness of the clustering performance. For all the metrics, higher value indicates better performance. More details of these metrics can be found in (Schütze, Manning, and Raghavan 2008).

Experimental Results

Experimental results on three public visual-tactile datasets are reported by comparing with the state-of-the-arts in this subsection. Due to the randomness of missing data generation, all experiments are repeated in ten times and reported with the mean value. Generally, the observations are sum-

	PHAC-2 Dataset		LMT Dataset		GelFabric Dataset	
Method	ACC(%)	NMI(%)	ACC(%)	NMI(%)	ACC(%)	NMI(%)
SC1	40.62±0.64	67.05±0.60	51.32±1.19	76.07±0.32	49.50±0.69	72.98±0.31
SC2	30.20±0.95	56.67±0.60	15.02±0.26	42.61±0.27	45.87±0.76	72.92±0.34
ConcatPCA	45.38±1.04	69.17±0.64	40.78±0.48	68.16±0.21	47.95±1.64	74.56±0.84
GLMSC	37.38±0.17	64.57±0.47	41.30±1.11	68.37±0.83	50.88±1.01	75.55±0.14
VTFC	51.41±0.63	70.85±0.32	43.94±0.16	51.03±0.22	55.72±1.04	74.76±0.38
IMG	37.90±0.92	49.79±0.14	41.66±1.68	67.45±0.93	37.39±2.10	66.06±0.48
GRMF	33.16±1.62	60.54±0.73	26.59±0.71	57.89±0.37	40.97±0.99	72.69±0.37
UEAF	40.56±0.06	63.20±0.39	47.78±0.19	74.09±0.60	51.26±0.05	72.36±0.72
OURS	**53.30±0.69**	**74.47±0.18**	**54.81±1.36**	**80.37±0.40**	**59.89±0.42**	**81.60±0.37**

Table 1: ACC and NMI performance on the three visual-tactile datasets, when the missing rate is set to be 0.1.
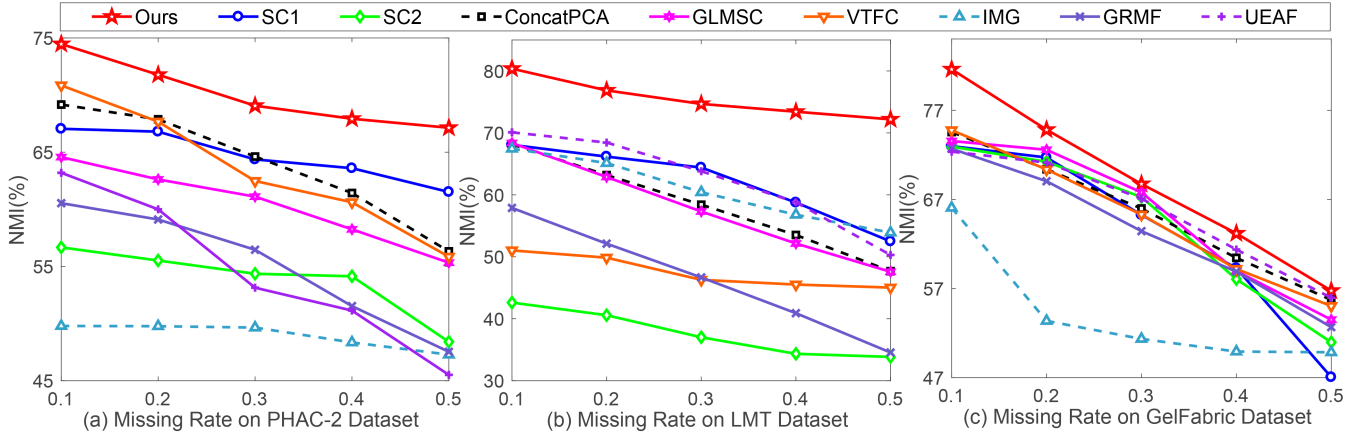

Figure 4: The average clustering NMI performance with respect to different missing rate on the (a) PHAC-2 dataset, (b) LMT dataset and (c) GelFabric dataset.
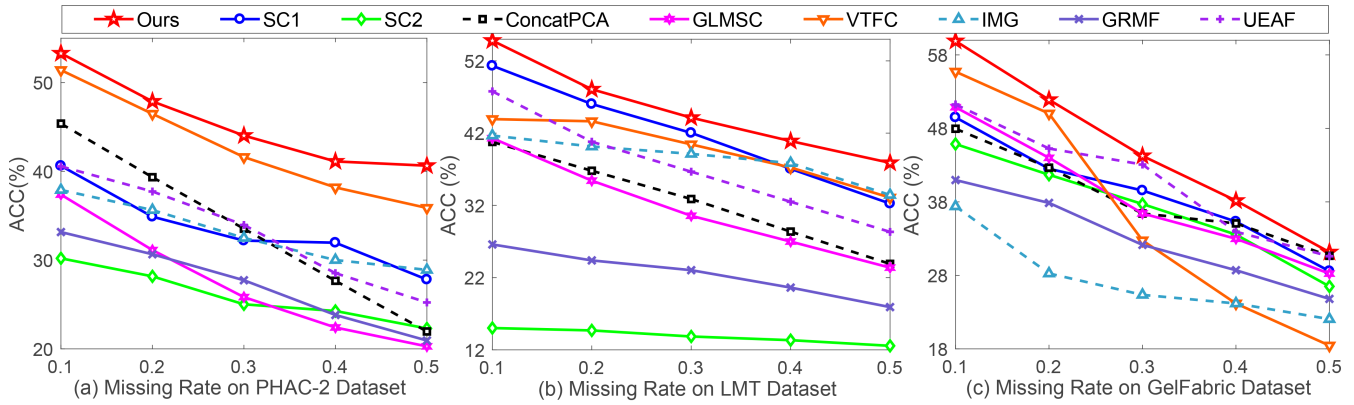

Figure 5: The average clustering ACC performance with respect to different missing rate on the (a) PHAC-2 dataset, (b) LMT dataset and (c) GelFabric dataset.

marized as follows: **1**) As shown in Table 1, where the missing rate is set to be 0.1, our GPVTF model consistently outperforms other methods with a clear improvement. For instance, compared with single-modality methods (*i.e.*, **SC1** and **SC2**), the performance is raised by 12.68% in ACC and 7.42% in NMI on the PHAC-2 dataset, which demonstrates the fact that fusing visual and tactile modalities does

improve the clustering performance. The results also show that our model is able to learn complementary information among the heterogeneous data. Compared with partial multi-view clustering method **UEAF** and visual-tactile fusing clustering method **VTFC**, the performance is raised by 1.89% and 3.62% in ACC and NMI, respectively. The reason why our GPVTF model achieves considerable achieve-
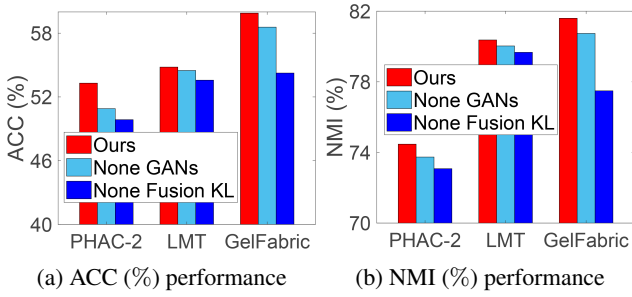
(a) ACC (%) performance (b) NMI (%) performance

Figure 6: Effectiveness of cross-modal clustering GANs, encoders, fusion KL-divergence losses, when the missing rate $\mathcal{MR}$ is 0.1.

(a) ACC (%) with different α (b) ACC (%) with different β

Figure 7: ACC (%) performance with different α and β when the missing rate $\mathcal{MR}$ is 0.1.

(a) NMI (%) performance   (b) ACC (%) performance

Figure 8: Performance with different φ_1 and φ_2 when the missing rate $\mathcal{MR}$ is 0.1.

ments is that our model can not only complete the missing data but also well align the heterogeneous data. **2**) As shown in Figure 3 and Figure 4, our GPVTF model outperforms other methods under different missing rates ($0.1 \sim 0.5$) on all the three datasets. Moreover, our model can also achieve competitive results on the PHAC-2 and LMT datasets even though the missing rate is very large. This observation indicates the effectiveness of the proposed conditional cross-modal clustering GANs. Besides, although the performance of **SC2** drops more slowly than ours, its performance is very low in most cases. We also find an interesting phenomenon that some multi-view clustering methods (*i.e.*, GRMF, IMG and GLMSC) even perform worse than single-view methods. The possible reason is that these methods do not take the gap between visual and tactile data into account. Directly fusion the heterogeneous data in a violent way would inevitably lead to performance degradation.

Ablation Study

The effect of the proposed cross-modal clustering GANs, fusion KL-divergence losses are analyzed first. Then we report the analyses of most important parameters α, β, φ_1 and φ_2.

Effectiveness of Cross-Modal Clustering GANs, Fusion KL-Divergence Losses: As shown in Figure 6, we first conduct ablation study to illustrate the effect of the proposed conditional cross-modal clustering GANs and fusion KL-Divergence losses when missing rate is set to be 0.1, where "None GANs" means the proposed conditional cross-modal clustering GANs are not employed and "None Fusion K-L" means the proposed fusion KL-Divergence losses are not employed, respectively. We can observe that "Ours" outperforms "None GANs" among all the datasets, which proves that the proposed conditional cross-modal clustering GANs promotes to achieve better performance. "Ours" outperforms than "None Fusion KL" proves that the proposed fusion KL-Divergence losses could better discover the information hidden in multi-modality data, and further enhance the performance.

Parameter Analysis: To explore the effect important weight coefficient α that controls the proportion of visual and tactile modalities, the parameter α is tuned from the set $\{0.1, 0.2, 0.3, 0.4, 0.5\}$, and report the clustering performance in Figure 7. Our model achieves the best clustering
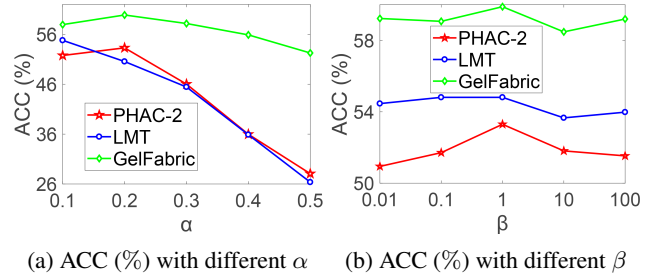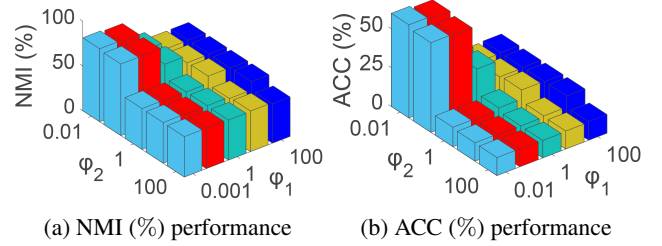
results, when the value of α is set to be 0.2, 0.2 and 0.1 on the PHAC-2, GelFabric and LMT datasets, respectively. Then, the parameter β is tuned from the set $\{0.01, 0.1, 1, 10, 100\}$, and the ACC performance is plotted in Figure 7. In fact, β controls the effect of common component, which further helps to update the encoders $E_1(\cdot)$ and $E_2(\cdot)$ simultaneously. It helps to ease the gap between visual and tactile modalities. It can be seen that when β is set to be 1, we gain the best performance. Thus we empirically choose $\beta = 1$ as default in this paper. To the end, we tune the trade-off parameters φ_1 and φ_2 in a similar way with β. As shown in Figure 8, our proposed GPVTF model performs best when φ_1 and φ_2 are set to be 0.01. Thus, we empirically choose $\beta = 1$, $\varphi_1 = 0.01$ and $\varphi_2 = 0.01$ as default in this paper in order to achieve the best performance.

Conclusion

In this paper, we put forward a Generative Partial Visual-Tactile Fused (GPVTF) framework, which tries to solve the problem of partial visual-tactile object clustering. GPVTF completes the partial visual-tactile data via two generators, which generate missing samples conditional on the other modality. In this way, the performance of clustering can be improved via the completed missing data and the aligned heterogeneous data. Moreover, pseudo-label based fusion KL-Divergence losses are leveraged to explicitly encapsulate the clustering task in our network, and further update the modality-specific encoders. Extensive experimental results on three public real-world benchmark visual-tactile datasets prove the superiority of our framework when comparing with several advanced methods.

References

Dang, Z.; Deng, C.; Yang, X.; and Huang, H. 2020. Multi-Scale Fusion Subspace Clustering Using Similarity Constraint. In *CVPR 2020*, 6658–6667.

Dong, J.; Cong, Y.; Sun, G.; and Hou, D. 2019. Semantic-Transferable Weakly-Supervised Endoscopic Lesions Segmentation. In *ICCV 2019*, 10711–10720.

Dong, J.; Cong, Y.; Sun, G.; Zhong, B.; and Xu, X. 2020. What Can Be Transferred: Unsupervised Domain Adaptation for Endoscopic Lesions Segmentation. In *CVPR 2020*, 4022–4031.

Gao, Y.; Hendricks, L. A.; Kuchenbecker, K. J.; and Darrell, T. 2016. Deep learning for tactile understanding from visual and haptic data. In *ICRA 2016*, 536–543. IEEE.

Hu, S.; Shi, Z.; and Ye, Y. 2020. DMIB: Dual-Correlated Multivariate Information Bottleneck for Multiview Clustering. *IEEE Transactions on Cybernetics* 1–15.

Hu, S.; Yan, X.; and Ye, Y. 2020. Dynamic auto-weighted multi-view co-clustering. *Pattern Recognition* 99.

Jiang, Y.; Xu, Q.; Yang, Z.; Cao, X.; and Huang, Q. 2019. DM2C: Deep Mixed-Modal Clustering. In *NeurlPS 2019*, 5880–5890.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *NeurlPS 2012*, 1097–1105.

Lee, J.; Bollegala, D.; and Luo, S. 2019. "Touching to See" and "Seeing to Feel": Robotic Cross-modal Sensory Data Generation for Visual-Tactile Perception. In *ICRA 2019*, 4276–4282. IEEE.

Li, S.-Y.; Jiang, Y.; and Zhou, Z.-H. 2014. Partial Multi-View Clustering. In *AAAI 2014*, 19681974. AAAI Press.

Li, Y.; Zhu, J.-Y.; Tedrake, R.; and Torralba, A. 2019. Connecting Touch and Vision via Cross-Modal Prediction. In *CVPR 2019*, 10609–10618.

Liu, H.; and Sun, F. 2018. *Robotic Tactile Perception and Understanding: A Sparse Coding Method*. Springer.

Liu, H.; Sun, F.; and Fang, B. 2019. Lifelong Learning for Heterogeneous Multi-Modal Tasks. In *ICRA 2019*, 6158–6164. IEEE.

Liu, H.; Yu, Y.; Sun, F.; and Gu, J. 2016. Visual–tactile fusion for object recognition. *IEEE Transactions on Automation Science and Engineering* 14(2): 996–1008.

Luo, S.; Yuan, W.; Adelson, E.; Cohn, A. G.; and Fuentes, R. 2018. ViTac: Feature Sharing Between Vision and Tactile Sensing for Cloth Texture Recognition. In *ICRA 2018*, 2722–2727. IEEE.

Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; and Chen, X. 2016. Improved techniques for training gans. In *NeurlPS 2016*, 2234–2242.

Schütze, H.; Manning, C. D.; and Raghavan, P. 2008. Introduction to information retrieval. In *Proceedings of the International Communication of Association for Computing Machinery Conference*, volume 4.

Shao, W.; Shi, X.; and Philip, S. Y. 2013. Clustering on multiple incomplete datasets via collective kernel learning. In *ICDM 2013*, 1181–1186. IEEE.

Strese, M.; Lee, J. Y.; Schuwerk, C.; Han, Q.; and Steinbach, E. 2014. A haptic texture database for tool-mediated texture recognition and classification. In *IEEE International Symposium on Haptic, Audio and Visual Environments and Games Proceedings*.

Strese, M.; Schuwerk, C.; and Steinbach, E. 2015. Surface classification using acceleration signals recorded during human freehand movement. In *IEEE World Haptics Conference*, 214–219. IEEE.

Sun, G.; Cong, Y.; Wang, Q.; Li, J.; and Fu, Y. 2020a. Lifelong Spectral Clustering. In *AAAI 2020*, 5867–5874. AAAI Press.

Sun, G.; Cong, Y.; Zhang, Y.; Zhao, G.; and Fu, Y. 2020b. Continual Multiview Task Learning via Deep Matrix Factorization. *IEEE Transactions on Neural Networks and Learning Systems* .

Wang, L.; Ding, Z.; Tao, Z.; Liu, Y.; and Fu, Y. 2019. Generative multi-view human action recognition. In *ICCV 2019*, 6212–6221.

Wang, Q.; Ding, Z.; Tao, Z.; Gao, Q.; and Fu, Y. 2018a. Partial multi-view clustering via consistent GAN. In *ICDM 2018*, 1290–1295.

Wang, Q.; Lian, H.; Gan, S.; Gao, Q.; and Jiao, L. 2020. iCmSC: Incomplete Cross-modal Subspace Clustering. *IEEE Transactions on Image Processing* 99(9): 1–11.

Wang, S.; Wu, J.; Sun, X.; Yuan, W.; Freeman, W. T.; Tenenbaum, J. B.; and Adelson, E. H. 2018b. 3d shape perception from monocular vision, touch, and shape priors. In *IROS 2018*, 1606–1613.

Wei, K.; Deng, C.; and Yang, X. 2020. Lifelong Zero-Shot Learning. In *IJCAI 2020*, 551–557. IJCAI Organization.

Wei, K.; Yang, M.; Wang, H.; Deng, C.; and Liu, X. 2019. Adversarial Fine-Grained Composition Learning for Unseen Attribute-Object Recognition. In *ICCV 2019*, 3741–3749.

Wen, J.; Zhang, Z.; Xu, Y.; Zhang, B.; Fei, L.; and Liu, H. 2019. Unified embedding alignment with missing views inferring for incomplete multi-view clustering. In *IJCAI 2019*.

Wen, J.; Zhang, Z.; Xu, Y.; and Zhong, Z. 2018. Incomplete multi-view clustering via graph regularized matrix factorization. In *ECCV 2018*, 0–0.

Xie, J.; Girshick, R.; and Farhadi, A. 2016. Unsupervised deep embedding for clustering analysis. In *ICML 2016*, 478–487.

Xu, C.; Guan, Z.; Zhao, W.; Wu, H.; Niu, Y.; and Ling, B. 2019. Adversarial incomplete multi-view clustering. In *IJCAI 2019*, 3933–3939. AAAI Press.

Yang, X.; Deng, C.; Wei, K.; Yan, J.; and Liu, W. 2020. Adversarial Learning for Robust Deep Clustering. *NeurlPS 2020* 33.

Yuan, W.; Wang, S.; Dong, S.; and Adelson, E. 2017. Connecting look and feel: Associating the visual and tactile properties of physical materials. In *CVPR 2017*, 5580–5588.

Zhang, C.; Fu, H.; Hu, Q.; Cao, X.; Xie, Y.; Tao, D.; and Xu, D. 2018. Generalized latent multi-view subspace clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence* .

Zhang, T.; Cong, Y.; Sun, G.; Wang, Q.; and Ding, Z. 2020. Visual Tactile Fusion Object Clustering. In *AAAI 2020*, 10426–10433. AAAI Press.

Zhao, H.; Liu, H.; and Fu, Y. 2016. Incomplete multi-modal visual data grouping. In *IJCAI 2016*, 2392–2398.

Zhao, Y.; Wang, Z.; and Huang, Z. 2021. Automatic Curriculum Learning With Over-repetition Penalty for Dialogue Policy Learning. In *AAAI 2021*. AAAI Press.

Zhao, Y.; Wang, Z.; Yin, K.; Zhang, R.; Huang, Z.; and Wang, P. 2020. Dynamic Reward-Based Dueling Deep Dyna-Q: Robust Policy Learning in Noisy Environments. In *AAAI 2020*, 9676–9684. AAAI Press.

Zheng, H.; Fang, L.; Ji, M.; Strese, M.; Özer, Y.; and Steinbach, E. 2016. Deep learning for surface material classification using haptic and visual information. *IEEE Transactions on Multimedia* 18(12): 2407–2416.