Few-Shot Audio-Visual Class-Incremental Learning with Temporal Prompting and Regularization

Yawen Cui[1], Li Liu[2*], Zitong Yu[3*], Guanjie Huang[2], Xiaopeng Hong[4]
[1] The Hong Kong Polytechnic University
[2] The Hong Kong University of Science and Technology (Guangzhou)
[3] Great Bay University
[4] Harbin Institute of Technology
yawencui@polyu.edu.hk, avrillliu@hkust-gz.edu.cn, yuzitong@gbu.edu.cn, ghuang565@connect.hkust-gz.edu.cn, hongxiaopeng@ieee.org

Abstract

Audio-Visual Learning (AVL) aims at the audio-visual perception with both audio and vision modalities. AVL also suffers from data insufficiency in many applications as with other unimodal tasks. Concurrently, AVL often needs to continuously learn over time rather than all knowledge simultaneously. Considering the above two perspectives, our work mainly focuses on benchmarking the unexplored Few-Shot Audio-Visual Class-Incremental Learning (FS-AVCIL), *i.e.*, continually perceiving novel categories described by a limited number of labeled examples with audio and visual modalities. Firstly, we provide the detailed task configuration together with a thorough analysis of the challenges in FS-AVCIL: (1) how to efficiently learn and fuse multimodal information with limited labeled examples; and (2) how to alleviate catastrophic forgetting cross-modal semantic correlations with limited data. Then, we propose an efficient framework based on Vision Transformer to solve FS-AVCIL, containing two parts: temporal-residual prompting for audio-visual synergy adapter and temporal prompt regularization. Specifically, temporal-residual prompting is incorporated into the audio-visual adapter to efficiently finetune the pre-trained foundation model with limited data and capture audio-visual correlation by learning temporal-relevant prompts. Besides, we regularize temporal-relevant prompts to memorize previous knowledge by fully using the temporal knowledge from various perspectives. This framework is validated in audio-visual classification tasks under the FS-AVCIL scenario, and extensive experiments demonstrate its superior performance.

Introduction

Data in visual tasks is of multi-modal characteristics, which inspires researchers to apply another one or two modalities (*e.g.*, audio and text) to compensate for information deficiencies and biases of the visual modality (Huang et al. 2021; Chen et al. 2020; Zhu et al. 2021). In this paper, we focus on audio-visual learning (AVL), which aims at audio-visual perception with both audio and vision modalities. The target of AVL is modeling audio-visual modalities jointly to capture cross-modal semantic correlations effectively. To achieve this, the main challenge is how to fully take advantage of multimodal information to acquire these correlations.

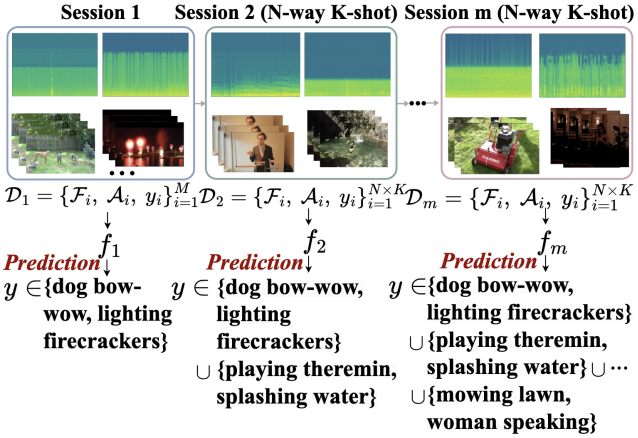

Figure 1: The configuration of FS-AVCIL. In the first session, a substantial dataset is presented. In subsequent sessions, the continual learning system is faced with a few-shot audio-visual learning challenge. Here we employ a 2-way 2-shot scenario as an of the N-way K-shot few-shot audio-visual learning problem.

Due to the modality discrepancy, learning-based AVL methods, aggregating multimodal information to obtain the semantic correlations generally requires encoding features from different modalities into a common latent space and then mapping the latent representations into the task space, which consists of complicated architectures. These fusion architectures usually contain numerous parameters trained on a large amount of accurately labeled data.

However, collecting a large number of samples is prohibited due to privacy (*e.g.*, medical images.) or even impossible due to the rarity (*e.g.*, rare species like pandas.). Moreover, annotating examples requires expert knowledge, and it is always costly. Few-Shot Learning (FSL) (Li, Wang, and Hu 2021; Finn, Abbeel, and Levine 2017; Snell, Swersky, and Zemel 2017; Zhang et al. 2024) is proposed to tackle this problem in which annotated samples are limited. As with unimodal vision tasks, AVL also suffers from data in-

*Li Liu and Zitong Yu are corresponding authors.

sufficiency in many applications, (*e.g.*, abnormal behavior detection in hospitals.). This data issue may be more frequent since collecting and annotating two types of data (*i.e.*, audio and video) nearly doubles the cost.

Humans can continuously obtain new knowledge without forgetting the old ones. Akin to human cognition, AVL systems also need to continuously perceive new concepts in real applications. For example, the abnormal behavior detection system should update continually to recognize novel behaviors. This capability is Class-Incremental Learning (CIL), which is one increment of continual learning (Gao et al. 2023; Ge et al. 2023; De Lange et al. 2021), and one of the well-known issues in CIL is catastrophic forgetting.

This paper conducts the pioneering research by considering these two capabilities (*i.e.*, FSL and CIL.) in AVL tasks simultaneously, which constitutes Few-Shot Audio-Visual Class-Incremental Learning (FS-AVCIL), *i.e.*, continually perceiving novel categories described by a limited number of labeled examples composing audio and video modalities, which is challenging and practical. This configuration simulates the neurological adaptive mechanisms of humans, which is illustrated in Fig 1.

As a task concluding multiple capabilities, the challenges of FS-AVCIL are two-fold: (1) *how to efficiently learn and fuse multimodal information with limited labeled examples to capture cross-modal semantic correlations ?* Learning with limited data faces the overfitting issue, and modeling the distribution of classes in the feature space is always difficult. For data with audio and video modalities, it is essential to alleviate the modality variance and efficiently obtain the audio-visual semantic correlation. (2) *How to alleviate catastrophic forgetting of cross-modal semantic semantic correlations in few-shot multimodal scenarios?* As incremental steps develop, the model tends to forget audio-visual correlations of previous tasks, resulting in performance degradation. Compared with unimodal continual learning, mitigating the forgetting issue is executed in a correlated perspective.

Current frameworks of audio-visual continual learning or few-shot audio-visual learning can only tackle one of these challenges. In this paper, we explore whether the audio-visual semantic correlations can benefit few-shot learning and continual learning simultaneously. First, by simply finetuning the model on new classes, the advantage of joint audio-visual modeling is obvious, and it achieves superior performance to that of the single audio or visual modality in a few-shot class-incremental paradigm. Inspired by the findings and to better model the audio-visual correlation under the limited data and continual learning regime, an efficient framework based on Vision Transformer (ViT) (Dosovitskiy et al. 2020) with two novel modules is proposed to solve the challenges of FS-AVCIL.

The framework is built with adapters (Houlsby et al. 2019; Jie and Deng 2022) which is one of efficient finetuning methods for large models. To make full use of limited multimodal data with potential knowledge, we propose an audio-visual synergy adapter with temporal-residual prompting. Specifically, in each adapter, audio features and video features are interactive for mitigating the modality variance and mapping to a shared space to better model the correlation. Then, inspired by the recent success of prompt learning in language models, temporal-residual prompting is assembled into each adapter to learn the temporal-relevant prompts from the two modalities, which is summed to a few random initialized tokens. Since the audio-visual correlation can benefit continual learning (Pian et al. 2023) and temporal information is beneficial for audio-visual perception (Huang et al. 2018), temporal prompt regularization is employed to memorize previous cross-modal semantic correlations by fully using the temporal knowledge from different perspectives.

We validate the effectiveness of our framework in audio-visual classification tasks and construct novel protocols with two benchmark datasets for FS-AVCIL. Our contributions are three-fold:

- To explore the efficacy of audio-visual correlation in the alleviation of forgetting and overfitting problems in few-shot class-incremental learning, we conduct the first work on few-shot audio-visual class-incremental learning, *i.e.*, perceiving audio and video modalities continually under the scenario of limited data regime.
- We propose an efficient learning framework for FS-AVCIL with temporal-residual prompting and audio-visual synergy adapters, which efficiently fine-tunes with limited data and learns temporal-relevant prompts.
- We leverage the temporal prompt regularization to force models to preserve previously learned audio-visual temporal correlations.
- We validate the efficacy of our framework on two audio-visual benchmark datasets with novel protocols.

Related Works

Few-Shot Learning Few-shot learning (FSL) is to identify new classes using only a limited number of support samples. Current FSL methods can be divided into three categories. (1) Data augmentation-based methods aim to expand the restricted labeled dataset by applying data augmentation (Li, Wang, and Hu 2021). (2) Optimization-based methods (Ravi and Larochelle 2016; Baik et al. 2021; Finn, Abbeel, and Levine 2017; Nichol, Achiam, and Schulman 2018) concentrate on creating a robust initialization or optimization strategy, enabling the model to quickly adapt to new tasks. (3) Metric-based methods (Snell, Swersky, and Zemel 2017; Sung et al. 2018; Afrasiyabi et al. 2022) learn a suitable distance function and make predictions based on the similarity between support and query samples. Audio-visual learning aims to perceive both visual and audio modalities. Most FSL algorithms focus solely on unimodal learning problems, and some of these studies conduct works on few-shot unimodal (video or audio) classification tasks (Cao et al. 2020; Yu et al. 2023; Wang and Anderson 2022; Liu, Zhang, and Pirsiavash 2023).

Continual Learning Recently, three main types of continual learning algorithms have emerged. (1) Regularization-based methods (Kirkpatrick et al. 2017; Aljundi et al. 2018;

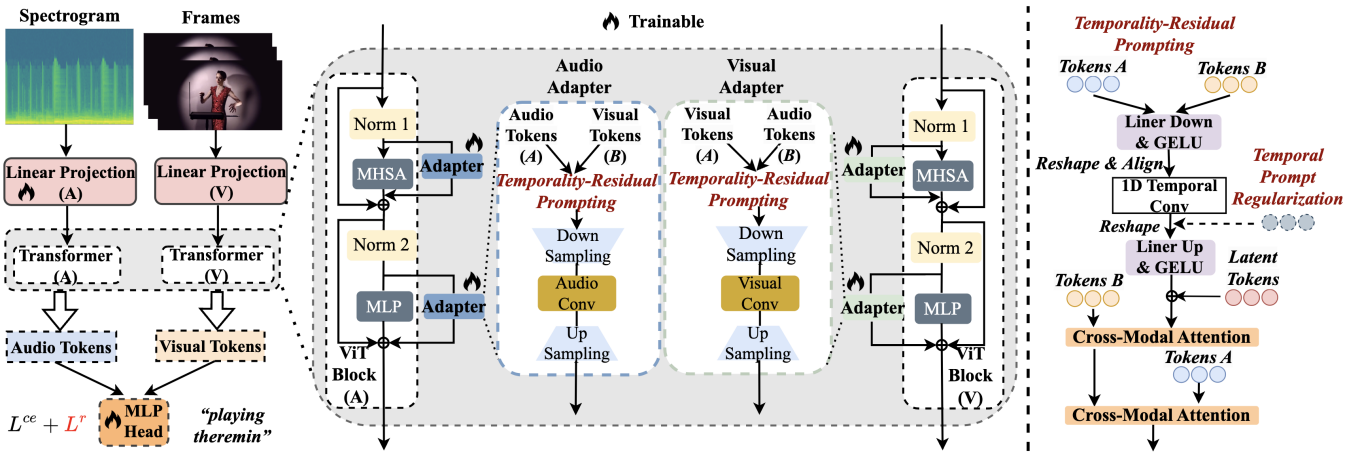

Figure 2: Overview of the framework. There are two kinds of audio-visual synergy adapters: (i) audio adapter, which incorporates visual features into the audio representation; and (ii) visual adapter, which injects audio features into the visual representation. The temporal-residual prompting for the audio-visual adapter is parallel with the MLP layers and MHSA layers. We combine the temporal prompt regularization loss into the final loss function for the memorizing purpose. Besides, the MLP head is expanded to adjust to the increase of novel classes. 'MHSA' and 'MLP' are short for the multi-head self-attention and multi-layer perceptron.

Li and Hoiem 2017; Yang et al. 2023) aim to prevent catastrophic forgetting by restricting the learning rate on parameters that are crucial for previous tasks. (2) Rehearsal-based methods (Rebuffi et al. 2017; Riemer et al. 2018; Hu et al. 2019) maintain a data buffer to store samples from older tasks, which are then used for training with data from the current task. (3) Distillation-based methods (Hou et al. 2019; Wu et al. 2019; Tao et al. 2020a; Guo et al. 2023) apply the technique of knowledge distillation to mitigate catastrophic forgetting. Continual learning under few-shot learning scenarios (Tao et al. 2020b; Dong et al. 2021) is also a popular topic in recent years.

Methodology

Problem Setup Firstly, a sequence of tasks, $\mathcal{T} = \{t_1, t_2, ..., t_n\}$, is usually provided, and the dataset sequence is $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, ..., \mathcal{D}_n\}$. Here $\mathcal{D}_1$ is the large-scale base dataset used in the first base session and the following datasets are all novel few-shot datasets, formulating as N-way-K-shot classification tasks, *i.e.,* each task includes N classes with K samples for each class. $\mathcal{C}_i$ represents the category set of session i. For $i \neq i'$, there is no overlap between the categories and samples of different sessions, *i.e.*, $\mathcal{C}_i \cap \mathcal{C}_{i'} = \emptyset$ and $\mathcal{D}_i \cap \mathcal{D}_{i'} = \emptyset$, where $i \neq i'$. We denote each audio-visual sample as $\mathcal{X} = \{\mathcal{F}, \mathcal{A}, y\}$. For the video modality $\mathcal{F}$, given a video clip of length t seconds, a temporal sequence of RGB frames $\mathcal{F} = \{\mathbf{V}_1, \mathbf{V}_2, ..., \mathbf{V}_F\}$ is uniformly sampled, where F is the temporal length. For the j-th frame in $\mathcal{F}$, $\mathbf{V}_j \in \mathbb{R}^{H \times W \times 3}$ with spatial dimensions $H \times W$. For audio modality $\mathcal{A}$, audio is transformed into a spectrogram $\mathcal{A} \in \mathbb{R}^{M \times C}$.

Overview of the Framework As illustrated in Fig 2, the tasks are executed by order in the sequence to mimic the process of novel tasks continually encountered. Our efficient learning framework for FS-AVCIL is a dual-stream structure built on ViT (Dosovitskiy et al. 2020). First, the audio frames and the video spectrogram are inputted into corresponding linear projections, which are then followed by L transformer blocks. As illustrated in Fig. 2, the audio-visual adapter is assembled into each transformer block as a parallel module of Multi-Layer Perception (MLP) and Multi-Head Self-Attention (MHSA). In each adapter, the temporal-residual prompting is incorporated before down-sampling to encode temporal information. Besides, temporal prompt regularization is conducted to memorize previous knowledge.

Temporal-Residual Prompting for Audio-Visual Synergy Adapter

Our efficient learning framework takes advantage of adapters, which is one of the efficient finetuning methods for large models. This structure enables the framework to adapt with limited labeled annotated data since fewer parameters are required to be fine-tuned to extend the recognition capability of novel categories. LAVISH (Lin et al. 2023) is the first work on efficient audio-visual learning based on adapters and proposes audio-visual adapters. Our audio-visual efficient learner is implemented based on LAVISH.

For video-based or audio-based missions, the temporal information can contribute to a better classifier (Wu et al. 2015; Cao et al. 2020). LAVISH executes with inputs of the single frame and the corresponding audio slice. It introduces randomly initialized latent tokens to compress and fuse audio and visual tokens, which do not consider temporal information during the multi-modal token compression and fusion. In this way, LAVISH is not suitable for audio-visual classification tasks since the video and the corresponding video are processed as a whole, *i.e.*, all frames extracted from the videos and the audio spectrogram are processed together. In order to efficiently integrate multi-

modal tokens and capture temporal audio-visual features for making full use of the multi-modal information, Temporal-Residual Prompting for the Audio-Visual synergy Adapter (TRP-AVA), placed in parallel with MHSA and MLP modules which are shown in Fig. 2, is proposed to capture the temporal information. In TRP-AVA, the audio-visual adapter is trainable to aggregate visual and audio tokens efficiently, while freezing all the pre-trained parameters of the transformer blocks.

Transformer Blocks assembled Audio-Visual synergy Adapter with Temporal-Residual Prompting. Our framework consists of two types of adapters: (i) audio adapter, which incorporates visual features into the audio representation; and (ii) visual adapter, which injects audio features into visual representation. The operation of TRP-AVA in transformer block l is denoted as $\mathcal{F}_l(\mathbf{T}_l^A, \mathbf{T}_l^B)$. Here we use $\mathbf{T}_l^A$ to represent tokens of the main modality and $\mathbf{T}_l^B$ to represent tokens of the incorporated modality. Then, the MHA operation in layer l can be written as:

$$\mathbf{T}_l = \mathbf{T}_l^A \ + \ \mathrm{MHA}(\mathbf{T}_l^A) \ + \ \mathcal{F}_l(\mathbf{T}_l^A, \ \mathbf{T}_l^B). \tag{1}$$

The MLP operation in layer l can be written as:

$$\mathbf{T}_l^A = \mathbf{T}_l \ + \ \mathrm{MLP}(\mathbf{T}_l) \ + \ \mathcal{F}_l(\mathbf{T}_l, \ \widehat{\mathbf{T}}_l), \tag{2}$$

where $\widehat{\mathbf{T}}^l$ is the parallel result of another type of adapter obtained by regarding $\mathbf{T}_l^B$ as the main modality.

Temporal-Residual Prompting. The inputs of each audio-visual synergy adapter are audio tokens and visual tokens, and here temporal-residual prompting illustrated in Fig. 2 (Right) can be regarded as the preliminary operation in each adapter to learn temporal-relevant prompts for the following cross-modal attention.

In TRP-AVA, $\mathbf{T}_l^A$ and $\mathbf{T}_l^B$ all go through the downsampling layer first. We denote the tokens after downsampling in a certain transformer layer as $\mathcal{T} = \{\overline{\mathbf{T}}_l^A, \ \overline{\mathbf{T}}_l^B\}$ with $\overline{\mathbf{T}}_l^A$ and $\overline{\mathbf{T}}_l^B$ as tokens of dominate modality and incorporated modality, respectively. After alignment and reshaping, the audio and visual tokens are concatenated first. Then, the 1D convolution operation with temporal kernel size k is applied to aggregate the local temporal information. We denote the obtained tokens as

$$\mathbf{T}_l^{A,B} = \mathrm{Conv1d}(\mathrm{Concat}(\overline{\mathbf{T}}_l^A, \ \overline{\mathbf{T}}_l^B)). \tag{3}$$

After the 1D convolution, $\mathbf{T}_l^{A,B}$ goes through the upsampling layer and GELU, and we obtain $\hat{\mathbf{T}}_l^{A,B}$.

Besides, a small set of randomly initialized latent tokens $\mathbf{L}_l$ is also introduced into the prompting, which is used to control the influence of temporal information on the model's final decision. Here we use cross-modal attention operation (Lin et al. 2023) defined as:

$$\mathrm{CMA}(\mathbf{Q}, \ \mathbf{K}, \ \mathbf{V}) = \mathbf{Q} \ + \ \mathrm{g} \cdot \mathrm{Softmax}(\mathbf{Q}\mathbf{K}^{\mathrm{T}})\mathbf{V}, \tag{4}$$

where g represents the learnable scalar parameter that regulates the transfer of information between different modalities, while $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ correspond to the query, key, and value tokens, respectively.

The key step of temporal-residual prompting is to add the tokens $\hat{\mathbf{T}}_l^{A,B}$ to latent tokens: $\hat{\mathbf{L}}_l = \mathbf{L}_l \ + \ \hat{\mathbf{T}}_l^{A,B}$. Then, the first cross-modal attention is executed to compress the tokens of incorporated modality: $\tilde{\mathbf{T}}_l^B = \mathrm{CMA}(\hat{\mathbf{L}}_l, \ \mathbf{T}_l^B, \ \mathbf{T}_l^B)$. In the end, another cross-modal attention is to aggregate the compressed tokens of the incorporated modality and the tokens of the main modality:

$$\tilde{\mathbf{T}}_l^{A,B} = \mathrm{CMA}(\mathbf{T}_l^A, \ \tilde{\mathbf{T}}_l^B, \ \tilde{\mathbf{T}}_l^B), \tag{5}$$

where $\tilde{\mathbf{T}}_l^{A,B}$ is a newly computed representation of the main modality by aggregating the cues of another modality.

Temporal Prompt Regularization

Another issue in audio-visual class-incremental learning (Pian et al. 2023) is the catastrophic forgetting of previous semantic correlations between audio and visual features as incremental step grows. In the limited annotated data regime, this issue is more severe since capturing the audio-visual correlation of novel classes always requires a large learning rate to update the parameters of aggregation modules, which is not beneficial for preserving previous knowledge. To alleviate the forgetting issue and with temporal-residual prompting, this paper proposes an exemplar-free module named Temporal Prompt Regularization (TPR). The exemplar-free manner does not require storing previous data, which can protect data privacy.

As temporal information is important in video understanding tasks (Huang et al. 2018), with the samples of the current session, TPR restricts the changes of temporal prompts obtained by the models in two neighboring sessions. As shown in Fig. 2, in terms of computation efficiency, TPR uses the temporal prompts before the linear up operation in each TRP-AVA. For the transformer block l, here we denote the temporal prompts obtained with the reference model $\mathcal{F}_{i-1}$ of session $i-1$ and the obtained temporal prompts of the current model $\mathcal{F}_i$ in session i as $\mathbf{T}_{i-1,l}^t$ and $\mathbf{T}_{i,l}^t$, respectively. Assuming that there are M temporal prompts in the transformer block l, the regularization loss is implemented with the cosine similarity loss, and computed along the prompt dimension and by averaging over the number of prompts. In session i, the regularization loss of the transformer block l is computed as

$$\mathcal{L}_{i,l}^r = \sum_{m=1}^{M} \left(1 \ - \ \frac{\mathbf{T}_{i-1,l,m}^t \cdot \mathbf{T}_{i,l,m}^t}{\left\| \mathbf{T}_{i-1,l,m}^t \right\| \left\| \mathbf{T}_{i,l,m}^t \right\|} \right), \tag{6}$$

where $\mathbf{T}_{i,l,m}^t$ and $\mathbf{T}_{i-1,l,m}^t$ represent the m^{th} token obtained in $\mathbf{T}_{i,l}^t$ and $\mathbf{T}_{i-1,l}^t$, respectively.

In session i, the total training loss is defined as

$$\mathcal{L}_i = \mathcal{L}_i^{ce} \ + \ \lambda \sum_{l=1}^{L} \mathcal{L}_{i,l}^r, \tag{7}$$

where $\mathcal{L}_i^{ce}$ is the cross-entropy loss and λ is the weight for the regularization loss. With the total training loss, the tasks are executed by session, and the multi-modal learning framework updates as novel tasks arrive.

	Method	Session ID 1	2	PD↓	Average Acc.
	Ft-ViT	69.29	18.12	51.17	43.71
Multimodal Learning	G-Blend	84.07	48.38	35.69	66.23
	MBT	82.73	45.60	37.13	64.17
	LAVISH	86.15	52.73	33.42	69.44
Continual Learning	LWF	69.29	22.41	46.88	45.85
	EWC	69.29	19.53	49.76	44.41
	iCaRL	69.29	31.12	38.07	38.17
	DER	69.10	23.91	45.19	46.51
	L2P	72.86	32.61	40.25	52.74
AV Continual Learning	AV-CIL	90.37	55.69	34.68	73.03
	CIGN	88.93	53.21	35.72	71.07
FS-AVCIL	**Ours**	90.71	**66.81**	**23.90**	**78.76**

Table 1: Results of protocol-1 on AVE dataset. The three values marked in blue correspond to three evaluation metrics.

	Method	Session ID 1	2	3	PD↓	Average Acc.
	Ft-ViT	69.29	17.05	15.15	54.14	33.83
Multimodal Learning	G-Blend	84.07	50.23	34.83	49.24	56.38
	MBT	82.73	48.07	31.94	50.79	54.25
	LAVISH	86.15	54.11	40.30	45.85	60.19
Continual Learning	LWF	69.29	20.07	17.44	51.85	35.60
	EWC	69.29	18.42	16.25	53.04	34.79
	iCaRL	69.29	33.08	28.88	40.41	43.75
	DER	69.10	25.54	17.22	51.88	37.29
	L2P	72.86	37.88	27.89	44.97	46.21
AV Continual Learning	AV-CIL	90.37	58.46	50.23	40.14	66.35
	CIGN	88.93	55.96	48.27	40.66	64.39
FS-AVCIL	**Ours**	90.71	66.79	**52.75**	**37.96**	**70.08**

Table 2: Comparative study of protocol-2 on AVE dataset.

Method	Session ID 1	2	3	4	5	PD↓	Average Acc.
Ft-ViT	69.29	30.26	13.58	8.04	7.96	61.33	25.83
G-Blend	84.07	63.26	44.37	34.05	33.57	50.5	51.86
MBT	82.73	61.31	40.25	30.98	30.07	52.66	49.07
LAVISH	86.15	65.89	47.95	37.57	36.57	49.58	54.83
LWF	69.29	31.17	16.30	15.95	12.40	57.29	29.02
EWC	69.29	29.38	15.90	15.03	11.98	57.31	28.32
iCaRL	69.29	44.72	33.08	27.19	26.14	43.15	40.08
DER	69.10	38.25	25.14	16.56	16.14	52.96	33.04
L2P	72.86	50.79	37.88	28.96	22.66	50.20	42.63
AV-CIL	90.37	69.90	58.76	51.25	48.63	41.74	63.78
CIGN	88.93	67.17	56.93	49.32	45.89	43.04	61.65
Ours	90.71	73.87	65.07	54.44	**50.51**	**40.20**	**66.92**

Table 3: Results of protocol-3 with AVE dataset.

Experiments

Protocols and Experimental Setup

AVE (Tian et al. 2018) dataset consists of events captured in 10-second video clips that feature both visual and auditory elements. It encompasses a total of 28 different event categories, with a collection of 4,143 videos. In our FS-AVCL setting, we sample 8 categories as base classes perceived in the first session, and the remaining 20 classes are encountered in the following incremental learning sessions. We use all the training samples of the 8 classes in the first session and the tasks of the incremental sessions are all under the few-shot learning scenario, *i.e.*, each novel class only contains limited annotated samples. We adopt comprehensive configurations of the incremental sessions: (1) 20-way 5-shot, (2) 10-way 5-shot, (3) 5-way 5-shot, and (3) 2-way 5-shot. The detailed protocol can be found in the Supplementary Materials. For the accumulative evaluation process, we use the original testing dataset.

Kinetics-Sounds is derived from the larger Kinetics-400 dataset (Kay et al. 2017). It consists of around 24,000 video clips, each with a duration of 10 seconds, and these clips are categorized into 32 classes of human actions. The dataset is divided into three parts: 20,000 videos for training, 2,000 for validation, and another 2,000 designated for testing. We initially selected 12 categories as the base classes for the first learning session. The subsequent incremental learning sessions then introduce the remaining 20 classes. We use all training samples from the 12 base classes during the first session. We implement a variety of configurations for the incremental learning sessions to thoroughly evaluate our approach. These configurations include: (1) 20-way 5-shot, (2) 10-way 5-shot, (3) 5-way 5-shot, and (3) 2-way 5-shot. The detailed protocol can be found in the Supplementary Materials. We employ the original testing set from the dataset for the cumulative performance evaluation.

Evaluation Metrics. Our evaluation is based on three evaluation metrics: (1) the final overall accuracy (%) obtained in the last session; (2) the Performance Dropping (PD) rate (%), which quantifies the absolute decline in accuracy in the last session compared to the first session; (3) the average accuracy (%) across all sessions.

Implementation Details. The framework was implemented with ViT (Dosovitskiy et al. 2020) as the backbone, and we loaded ViT-Base model pretrained on ImageNet-21K (Deng et al. 2009). For each video, we sampled 8 frames and these frames were introduced to the visual branch. During training, we froze all the parameters of 12 transformer blocks except for the Linear Projection of the audio branch and MLP head. The dimension of tokens is 768, and the number of latent tokens is 2. The proposed TRP-AVA is parallel with the MLP layer and MHSA in each transformer block. Besides, the downsampling dimension in TRP-AVA is 8, and the temporal kernel size k in 1D convolution operation is 5. For TPR, we use the temporal prompts of last 6 transformer blocks and the regulization loss weight λ is 0.3. Besides, we used Adam optimizer, and the model was trained for 30 epochs with a batch size of 2 (*i.e.*, 2 videos and 8 frames each.) and a learning rate of 0.0003. Our framework was implemented using Pytorch, and all experiments were conducted on NVIDIA A100 GPU.

	Method	Session ID											PD↓	Average Acc.
		1	2	3	4	5	6	7	8	9	10	11		
	Ft-ViT	69.29	40.69	22.06	13.00	10.47	13.36	8.06	7.83	5.52	6.72	6.72	50.77	18.52
Multimodal Learning	G-Blend	84.07	70.30	59.98	53.06	49.35	44.28	39.37	33.07	30.79	28.90	24.31	59.76	47.04
	MBT	82.73	68.51	56.73	50.49	45.21	40.72	36.58	29.50	24.16	23.02	20.15	63.58	43.44
	LAVISH	90.71	73.99	63.73	57.85	54.65	49.32	44.19	40.36	39.83	37.10	34.58	56.13	53.30
Continual Learning	LWF	69.29	41.62	25.21	23.14	15.76	16.24	14.93	11.63	11.52	12.72	12.97	56.32	23.18
	EWC	69.29	38.25	22.18	21.33	13.26	12.24	11.53	9.28	10.37	10.58	11.74	57.55	20.91
	iCaRL	69.29	50.29	40.69	37.67	29.84	27.74	22.90	22.29	20.35	21.24	19.65	49.64	32.90
	DER	69.10	28.53	18.93	16.53	12.90	12.84	11.16	10.92	10.91	10.18	11.21	57.89	19.38
	L2P	72.86	48.27	42.51	39.01	32.95	26.03	24.84	20.54	18.43	15.11	16.72	56.14	32.48
AV Continual Learning	AV-CIL	90.37	70.18	64.27	58.92	54.70	53.17	51.13	47.92	46.92	47.21	45.52	44.85	57.30
	CIGN	88.75	67.28	61.39	56.21	54.82	50.19	49.27	47.60	44.45	44.28	43.87	44.88	55.28
FS-AVCIL	**Ours**	90.71	75.28	68.35	66.70	63.53	60.62	55.48	49.10	48.02	49.40	**48.77**	**41.94**	**61.45**

Table 4: Results of protocol-4 on AVE dataset.

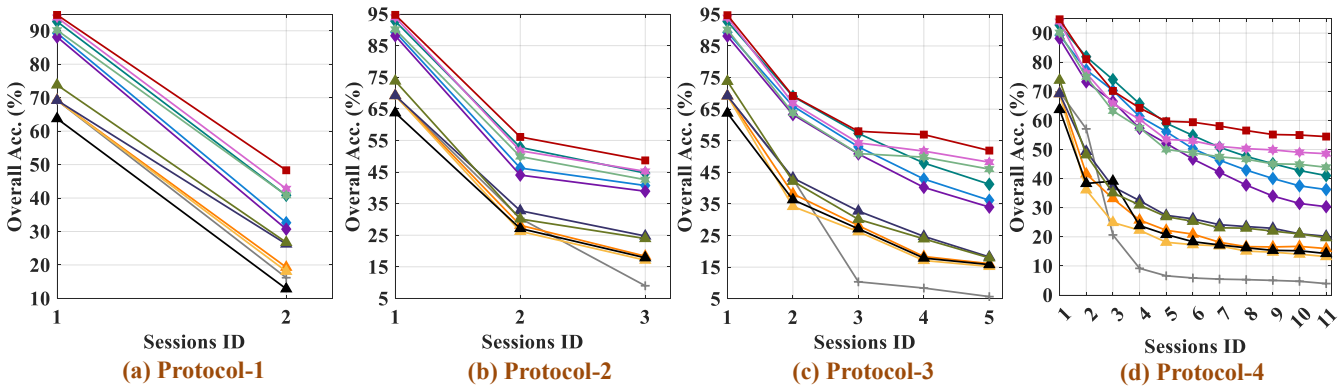

Figure 3: Comparative studies on four protocols with Kinetics-Sounds dataset.

Comparative Studies

We first provide baselines and comprehensive comparative studies from three perspectives. (1) Finetuning ViT (Ft-ViT). It means simply fine-tuning ViT with a few training samples of new classes, which is the lower bound of continual learning performance. (2) Multimodal learning methods (Wang, Tran, and Feiszli 2020; Nagrani et al. 2021; Lin et al. 2023). We incorporate these multimodal learning methods to fuse the information in a dual-stream of ViT-based framework. (3) Continual learning methods (Li and Hoiem 2017; Kirkpatrick et al. 2017; Rebuffi et al. 2017; Buzzega et al. 2020; Wang et al. 2022). With the dual-stream of ViT-based framework, we use a simple fusion by averaging outputs of two streams and apply these comparative methods for memorizing old knowledge. (4) Audio-visual continual learning methods (Pian et al. 2023; Mo, Pian, and Tian 2023). These methods also perform general class-incremental learning tasks, concluding the fusion and memorizing modules. We use these methods to execute the FS-AVCIL task.

Results on AVE dataset. The comparative study results are illustrated in Tab. 1, Tab. 2, Tab. 3, and Tab. 4, respectively. We analyze the results from the following three perspectives. (1) Compared with other multimodal learning methods, our method achieves remarkable performance in the final overall classification accuracy and the average accuracy and suffers from a lower performance dropping rate. This achievement demonstrates that TRP incorporating the temporal information of video can assist in effectively fusing the multi-modality by utilizing important information, thus capturing the audio-visual semantic correlation. This audio-visual semantic correlation can promote the final classification performance. (2) We compare our method with the other five continual learning methods. Our method achieves the surpassing performance toward three evaluation metrics. This phenomenon verifies that our TPR can make full use of the temporal representations of different blocks to alleviate the forgetting issue. Moreover, it also reflects that our proposed method is suitable for multi-modal modeling. (3) Our method is compared with two audio-visual continual learning methods proposed for general continual learning scenarios without considering the overfitting issue encountered in FSL. We can conclude that temporal-residual prompting for audio-visual learning is an efficient module for ViT that can adapt to new tasks with limited annotated data.

Results of Kinetics-Sound. The comparative study results are presented in Fig. 3. The curve obtained by our methods lies above all other curves, demonstrating that our method exceeds other comparative approaches. Compared with fine-

Modality	Session ID					PD↓	Average Acc.
	1	2	3	4	5		
Video	80.00	59.53	45.55	38.91	33.45	46.55	51.49
Audio	77.86	56.92	41.01	36.95	31.71	46.15	48.89
Video & Audio	90.71	73.87	65.07	54.44	**50.51**	**40.20**	**66.92**

Table 5: Unimodal learning with protocol-3 on AVE dataset.

tuning ViT, our method outperforms it by a large margin, illustrating the efficacy of our method in tackling the FS-AVCIL task. Furthermore, it is worth noting that the curve representing our method experiences a gentle downtrend, which represents that our method can better alleviate the overfitting issue resulting in less performance dropping. When the task sequence becomes longer (*e.g.*, protocol-3 and protocol-4), the performance suffers from fluctuation, not only the downtrend. This may result from the modality bias that is often encountered in multimodal learning. When the modality bias occurs, the dominant modality affects the learning status of another modality, thus leading to hard convergence and bad performance.

Further Discussions on the Uni-Modal Learning

We implemented few-shot unimodal class-incremental learning (*i.e.*, audio classification and video classification.) with the vanilla adapter-assembled ViT framework and temporal prompt regularization. As shown in Tab. 5, our framework for FS-AVCIL outperforms the vanilla adapter-assembled ViT framework of unimodal learning towards three evaluation indicators. This phenomenon demonstrates: (1) the proposed TRP-AVA assembled into each vision transformer block can achieve effective fusion and capture the audio-visual semantic correlation with limited annotated data efficiently; (2) The proposed TPR can better memorize this semantic correlation and it can better alleviate the forgetting of previous knowledge than unimodal learning.

Ablation Studies

Efficacy of TRP-AVA. We present comparative results with the configuration of removing TRP-AVA in each transformer block in Tab. 6. There is a huge performance gap between the results obtained with the framework with TRP-AVA and without TRP-AVA. This phenomenon results from the efficient fine-tuning strategy of TRP-AVA. With the audio-visual adapter assembled in ViT, We first transfer the knowledge of pretrained ViT to our audio-visual classification task, while leaving the parameters of TRP-AVA trainable to capture the specific knowledge of our task. In TRP-AVA, the temporal-residual prompting encodes temporal information to assist the aggregation process for obtaining multi-modal semantic correlations for specific classes.

Efficacy of TPR. The ablation results in Tab. 6 show that the model suffers from a severe performance drop due to the lack of a strategy for alleviating catastrophic forgetting. Compared with other methods of continual learning, TPR first proposes to regulate the obtained shared temporal prompt of audio and video and the regularization loss can

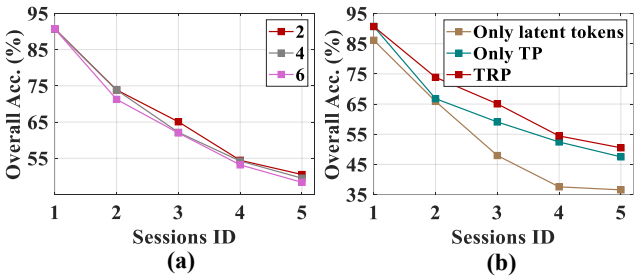

Figure 4: Ablation on (a) the number of tokens and (b) the residual operation in TRP-AVA. TP means only temporal tokens are used.

TRP-AVA	TPR	Session ID					PD↓	Average Acc.
		1	2	3	4	5		
✗	✗	69.29	30.26	13.58	8.04	7.96	61.33	25.83
✗	✓	69.29	40.21	34.75	26.28	22.60	46.69	38.63
✓	✗	90.71	65.89	47.95	37.57	36.57	54.14	55.74
✓	✓	90.71	73.87	65.07	54.44	**50.51**	**40.20**	**66.92**

Table 6: Ablation on modules with protocol-3 of AVE.

better maintain this temporal information, which is significant for audio-visual perception.

Impacts of the number of latent tokens. Two latent tokens with a small number of trainable parameters are adopted as the default setting due to the parameter-efficient purpose. We present the results with 4 and 6 tokens in Fig. 4 (a). With 2 tokens, the performance of FS-AVCIL tasks can take more advantage of our framework. This is because of the learning setting with limited labeled data. Overfitting, as one of the longstanding and widely-acknowledged issues that occurred in few-shot learning, would become severe when there are more trainable parameters in frameworks. Hence, 2 latent tokens with a few extra trainable parameters are suitable for our FS-AVCIL configuration.

Impacts of the residual operation in TRP-AVA. There is the residual operation on latent tokens by adding the tokens obtained by temporal-residual prompting, contributing to the best performance shown in Fig. 4 (b). In temporal-residual prompting, we capture the temporal information with audio and video tokens, and latent tokens are introduced to balance the contribution of temporal information and spatial information to final multi-modal decisions.

Conclusion

This paper conducts Few-Shot Audio-Visual Class-Incremental Learning (FS-AVCIL) for audio-visual classification tasks, which involves continuously recognizing new categories using a small set of labeled examples. We innovatively introduce two key components: TRP-AVA and TPR. Extensive experiments on novel benchmarks confirm its effectiveness.The former one enables adapters to fine-tune effectively with limited data while capturing temporal dynamics, and the latter one preserves previously learned information.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 62471420, 62101351, 62306061, 62076195, 62376070), and Guangdong Basic and Applied Basic Research Foundation (Grant No. 2023A1515140037).

References

Afrasiyabi, A.; Larochelle, H.; Lalonde, J.-F.; and Gagné, C. 2022. Matching feature sets for few-shot image classification. In *CVPR*.

Aljundi, R.; Babiloni, F.; Elhoseiny, M.; Rohrbach, M.; and Tuytelaars, T. 2018. Memory aware synapses: Learning what (not) to forget. In *ECCV*.

Baik, S.; Choi, J.; Kim, H.; Cho, D.; Min, J.; and Lee, K. M. 2021. Meta-learning with task-adaptive loss function for few-shot learning. In *ICCV*.

Buzzega, P.; Boschini, M.; Porrello, A.; Abati, D.; and Calderara, S. 2020. Dark experience for general continual learning: a strong, simple baseline. *NeurIPS*.

Cao, K.; Ji, J.; Cao, Z.; Chang, C.-Y.; and Niebles, J. C. 2020. Few-shot video classification via temporal alignment. In *CVPR*.

Chen, H.; Xie, W.; Vedaldi, A.; and Zisserman, A. 2020. Vggsound: A large-scale audio-visual dataset. In *ICASSP*.

De Lange, M.; Aljundi, R.; Masana, M.; Parisot, S.; Jia, X.; Leonardis, A.; Slabaugh, G.; and Tuytelaars, T. 2021. A continual learning survey: Defying forgetting in classification tasks. *IEEE TPAMI*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A large-scale hierarchical image database. In *CVPR*.

Dong, S.; Hong, X.; Tao, X.; Chang, X.; Wei, X.; and Gong, Y. 2021. Few-shot class-incremental learning via relation knowledge distillation. In *AAAI*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*.

Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*.

Gao, Q.; Zhao, C.; Sun, Y.; Xi, T.; Zhang, G.; Ghanem, B.; and Zhang, J. 2023. A Unified Continual Learning Framework with General Parameter-Efficient Tuning. In *ICCV*.

Ge, Y.; Li, Y.; Ni, S.; Zhao, J.; Yang, M.-H.; and Itti, L. 2023. CLR: Channel-wise Lightweight Reprogramming for Continual Learning. In *ICCV*, 18798–18808.

Guo, G.; Han, L.; Wang, L.; Zhang, D.; and Han, J. 2023. Semantic-aware knowledge distillation with parameter-free feature uniformization. *Visual Intelligence*, 1(1): 6.

Hou, S.; Pan, X.; Loy, C. C.; Wang, Z.; and Lin, D. 2019. Learning a unified classifier incrementally via rebalancing. In *CVPR*.

Houlsby, N.; Giurgiu, A.; Jastrzebski, S.; Morrone, B.; De Laroussilhe, Q.; Gesmundo, A.; Attariyan, M.; and Gelly, S. 2019. Parameter-efficient transfer learning for NLP. In *ICML*.

Hu, W.; Lin, Z.; Liu, B.; Tao, C.; Tao, Z. T.; Zhao, D.; Ma, J.; and Yan, R. 2019. Overcoming catastrophic forgetting for continual learning via model adaptation. In *ICLR*.

Huang, D.-A.; Ramanathan, V.; Mahajan, D.; Torresani, L.; Paluri, M.; Fei-Fei, L.; and Niebles, J. C. 2018. What makes a video a video: Analyzing temporal information in video understanding models and datasets. In *CVPR*.

Huang, Y.; Du, C.; Xue, Z.; Chen, X.; Zhao, H.; and Huang, L. 2021. What makes multi-modal learning better than single (provably). *NeurIPS*.

Jie, S.; and Deng, Z.-H. 2022. Convolutional bypasses are better vision transformer adapters. *arXiv preprint arXiv:2207.07039*.

Kay, W.; Carreira, J.; Simonyan, K.; Zhang, B.; Hillier, C.; Vijayanarasimhan, S.; Viola, F.; Green, T.; Back, T.; Natsev, P.; et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

Kirkpatrick, J.; Pascanu, R.; Rabinowitz, N.; Veness, J.; Desjardins, G.; Rusu, A. A.; Milan, K.; Quan, J.; Ramalho, T.; Grabska-Barwinska, A.; et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13): 3521–3526.

Li, J.; Wang, Z.; and Hu, X. 2021. Learning intact features by erasing-inpainting for few-shot classification. In *AAAI*.

Li, Z.; and Hoiem, D. 2017. Learning without forgetting. *IEEE TPAMI*.

Lin, Y.-B.; Sung, Y.-L.; Lei, J.; Bansal, M.; and Bertasius, G. 2023. Vision transformers are parameter-efficient audio-visual learners. In *CVPR*.

Liu, X.; Zhang, H.; and Pirsiavash, H. 2023. MASTAF: A Model-Agnostic Spatio-Temporal Attention Fusion Network for Few-shot Video Classification. In *WACV*.

Mo, S.; Pian, W.; and Tian, Y. 2023. Class-incremental grouping network for continual audio-visual learning. In *ICCV*.

Nagrani, A.; Yang, S.; Arnab, A.; Jansen, A.; Schmid, C.; and Sun, C. 2021. Attention bottlenecks for multimodal fusion. *NeurIPS*.

Nichol, A.; Achiam, J.; and Schulman, J. 2018. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*.

Pian, W.; Mo, S.; Guo, Y.; and Tian, Y. 2023. Audio-visual class-incremental learning. In *ICCV*.

Ravi, S.; and Larochelle, H. 2016. Optimization as a model for few-shot learning. In *ICLR*.

Rebuffi, S.-A.; Kolesnikov, A.; Sperl, G.; and Lampert, C. H. 2017. icarl: Incremental classifier and representation learning. In *CVPR*.

Riemer, M.; Cases, I.; Ajemian, R.; Liu, M.; Rish, I.; Tu, Y.; and Tesauro, G. 2018. Learning to Learn without Forgetting by Maximizing Transfer and Minimizing Interference. In *ICLR*.

Snell, J.; Swersky, K.; and Zemel, R. 2017. Prototypical networks for few-shot learning. *NeurIPS*.

Sung, F.; Yang, Y.; Zhang, L.; Xiang, T.; Torr, P. H.; and Hospedales, T. M. 2018. Learning to compare: Relation network for few-shot learning. In *CVPR*.

Tao, X.; Chang, X.; Hong, X.; Wei, X.; and Gong, Y. 2020a. Topology-preserving class-incremental learning. In *ECCV*.

Tao, X.; Hong, X.; Chang, X.; Dong, S.; Wei, X.; and Gong, Y. 2020b. Few-shot class-incremental learning. In *CVPR*.

Tian, Y.; Shi, J.; Li, B.; Duan, Z.; and Xu, C. 2018. Audio-visual event localization in unconstrained videos. In *ECCV*.

Wang, W.; Tran, D.; and Feiszli, M. 2020. What makes training multi-modal classification networks hard? In *CVPR*.

Wang, Y.; and Anderson, D. V. 2022. Hybrid attention-based prototypical networks for few-shot sound classification. In *ICASSP*.

Wang, Z.; Zhang, Z.; Lee, C.-Y.; Zhang, H.; Sun, R.; Ren, X.; Su, G.; Perot, V.; Dy, J.; and Pfister, T. 2022. Learning to prompt for continual learning. In *CVPR*.

Wu, Y.; Chen, Y.; Wang, L.; Ye, Y.; Liu, Z.; Guo, Y.; and Fu, Y. 2019. Large scale incremental learning. In *CVPR*.

Wu, Z.; Wang, X.; Jiang, Y.-G.; Ye, H.; and Xue, X. 2015. Modeling spatial-temporal clues in a hybrid deep learning framework for video classification. In *ACM MM*.

Yang, Y.; Cui, Z.; Xu, J.; Zhong, C.; Zheng, W.-S.; and Wang, R. 2023. Continual learning with Bayesian model based on a fixed pre-trained feature extractor. *Visual Intelligence*, 1(1): 5.

Yu, Z.; Wang, S.; Chen, L.; and Cheng, Z. 2023. Halluaudio: Hallucinate Frequency as Concepts For Few-Shot Audio Classification. In *ICASSP*.

Zhang, Z.; Yuan, M.; Ma, X.; Liu, Y.; Lu, H.; Wang, L.; Su, Y.; and Liu, Y. 2024. Unified Regularity Measures for Sample-wise Learning and Generalization. *Visual Intelligence*, 2(36).

Zhu, H.; Luo, M.-D.; Wang, R.; Zheng, A.-H.; and He, R. 2021. Deep audio-visual learning: A survey. *International Journal of Automation and Computing*, 18(3): 351–376.