Class-Incremental Grouping Network for Continual Audio-Visual Learning

Shentong Mo[†]
Carnegie Mellon University

Weiguo Pian[†]
University of Texas at Dallas

Yapeng Tian[*]
University of Texas at Dallas

Abstract

Continual learning is a challenging problem in which models need to be trained on non-stationary data across sequential tasks for class-incremental learning. While previous methods have focused on using either regularization or rehearsal-based frameworks to alleviate catastrophic forgetting in image classification, they are limited to a single modality and cannot learn compact class-aware cross-modal representations for continual audio-visual learning. To address this gap, we propose a novel class-incremental grouping network (CIGN) that can learn category-wise semantic features to achieve continual audio-visual learning. Our CIGN leverages learnable audio-visual class tokens and audio-visual grouping to continually aggregate class-aware features. Additionally, it utilizes class tokens distillation and continual grouping to prevent forgetting parameters learned from previous tasks, thereby improving the model's ability to capture discriminative audio-visual categories. We conduct extensive experiments on VGGSound-Instruments, VGGSound-100, and VGG-Sound Sources benchmarks. Our experimental results demonstrate that the CIGN achieves state-of-the-art audio-visual class-incremental learning performance. Code is available at https://github.com/stoneMo/CIGN.

1. Introduction

The strong correspondence between audio signals and visual objects in the world enables humans to perceive the source of a sound, such as a meowing cat. This perception intelligence has motivated researchers to explore audio-visual joint learning for various tasks, such as audio-visual event classification [60], sound source separation [69, 17, 58], and visual sound localization [55, 12, 39]. In this work, we focus on the problem of classifying sound sources from both video frames and audio in a continual learning [28, 37] setting, *i.e.*, train audio-visual learning models on non-stationary audio-visual pairs, enabling the model to classify sound sources in videos incrementally.

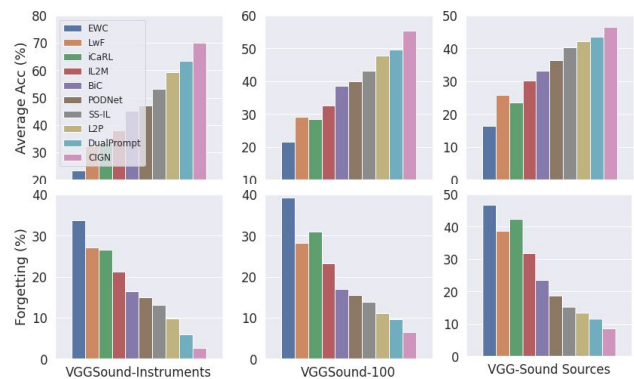[*]Corresponding author, [†]Equal contribution.

Figure 1. Comparison of our CIGN with state-of-the-art approaches on Average Accuracy (Top Row, higher is better) and Forgetting (Bottom Row, lower is better) for continual audio-visual learning on VGGSound-Instruments [26], VGGSound-100 [13], and VGG-Sound Sources [13] benchmarks.

Continual learning is a recently popular and challenging problem that aims to train models on non-stationary data given sequential tasks for class-incremental learning. Previous methods [28, 53, 34, 8, 65, 23, 51] mainly used either regularization or rehearsal-based approaches to alleviate catastrophic forgetting between old and new classes for image classification. A typical regularization-based work, EWC [28] addressed catastrophic forgetting in neural networks by training sequential models to protect the weights crucial for previous tasks. Similarly, LwF [34] used only new task data for training while preserving the original capabilities. To improve the performance under challenging datasets, BiC [65] adopted two bias parameters in a linear model to address the bias to new classes. SS-IL [2] combined separated softmax with task-wise knowledge distillation to solve this bias. However, those baselines are based on a single input modality and perform worse for continual audio-visual learning. In this work, we address a new multi-modal continual problem by extracting disentangled and compact representations with learnable audio-visual class-incremental tokens as guidance for continual learning.

Since prompts can learn task-specific knowledge [50, 31, 33, 25] for transfer learning, recent researchers have tried

to adapt diverse prompts-based pipelines to resolve continual learning challenges. The basic idea of prompting is to design a function to alter the input text for pre-trained large language models such that the model can generate more informative features about the new task. For example, L2P [63] leveraged a prompt pool memory space to instruct the model prediction and explicitly maintain model plasticity for task-invariant and task-specific knowledge. More recently, DualPrompt [62] proposed complementary prompts to instruct the pre-trained backbone for learning tasks arriving sequentially. Despite their promising results, they can only deal with one modality, and the design of audio-visual prompts requires choreographed heuristics. When we apply their prompting to our audio-visual settings, they cannot learn compact class-aware cross-modal representations for class-incremental learning.

The main challenge is that sounds are naturally mixed in the audio space such that the global audio representation extracted from the sound is easy to be catastrophically forgotten by the cross-modal model. This inspires us to disentangle the individual semantics for old and current audio-visual pairs to guide continual audio-visual learning. To address the problem, our key idea is to disentangle individual audio-visual representations from sequential tasks using audio-visual continual grouping for class-incremental learning, which differs from existing regularization-based and rehearsal-based methods. During training, we aim to learn audio-visual class tokens to continually aggregate category-aware source features from the sound and the image, where separated high-level semantics for sequential audio-visual tasks are learned.

To this end, we propose a novel class-incremental grouping network, namely CIGN, that can directly learn category-wise semantic features to achieve continual audio-visual learning. Specifically, our CIGN leverages learnable audio-visual class tokens and audio-visual grouping to continually aggregate class-aware features. Furthermore, it leverages audio-visual class tokens distillation and continual grouping to alleviate forgetting parameters learned from previous tasks for capturing discriminative audio-visual categories.

Empirical experiments on VGGSound-Instruments, VGGSound-100, and VGG-Sound Sources benchmarks comprehensively demonstrate the state-of-the-art performance against previous continual learning baselines, as shown in Figure 1. In addition, qualitative visualizations of learned audio-visual embeddings vividly showcase the effectiveness of our CIGN in aggregating class-aware features to avoid cross-modal catastrophic forgetting. Extensive ablation studies also validate the importance of class-token distillation and continual grouping in learning compact representations for class-incremental learning.

Our contributions can be summarized as follows:

- We present a novel class-incremental grouping network, namely CIGN, that can directly learn category-wise semantic features to achieve continual audio-visual learning.

- We introduce learnable audio-visual class tokens distillation and continual grouping to continually aggregate class-aware features for alleviating cross-modal catastrophic forgetting.

- Extensive experiments can demonstrate the state-of-the-art superiority of our CIGN over previous baselines on audio-visual class-incremental scenarios.

2. Related Work

Audio-Visual Learning. Audio-visual learning has been explored in many previous methods [5, 49, 4, 29, 55, 69, 68, 16, 56, 57, 45, 46, 48, 41, 40, 32] to capture audio-visual alignment from those two different modalities in videos. A comprehensive survey on audio-visual learning can be found in [64]. Given video sequences with audio and frames, the objective is to push away embeddings from non-matching audio-visual pairs while closing audio and visual representations from the matching pair. Such cross-modal alignment is helpful for several tasks, such as audio spatialization [47, 18, 11, 45], audio/speech separation [16, 17, 19, 69, 68, 21, 58, 20], visual sound source localization [55, 54, 24, 1, 52, 12, 39, 38, 43, 44], and audio-visual event parsing [60, 59, 66, 35, 42]. In this work, our main focus is to learn compact audio-visual representations on non-stationary audio-visual pairs given sequential tasks for class-incremental learning, which is more challenging than the abovementioned tasks on independent and identically distributed audio-visual data.

Continual Learning. Continual learning aims to train a single model on non-stationary data distributions where distinct classification tasks are given sequentially. Early works [28, 34, 3] mainly applied regularization on the learning rate on crucial parameters for old tasks to preserve the model capability. In recent years, diverse rehearsal-based pipelines [53, 8, 6, 23, 10, 65, 51, 7, 9, 2] have been proposed to resolve the catastrophic forgetting problem in challenging settings. A dual memory network was proposed in IL2M [6] to use both a bounded memory of the past images and a second memory for past class statistics obtained from initial training. By carefully balancing the compromise between old and new classes, PODNet [15] introduced a spatial-based distillation loss to optimize the model with multiple proxy embeddings for each category. Unlike them, we address the cross-modal catastrophic forgetting problem in audio-visual settings. Instead, we leverage the category-aware representations of individual audio-visual pairs to predict the corresponding category for sequential classification tasks, where learnable audio-visual class tokens are uti-
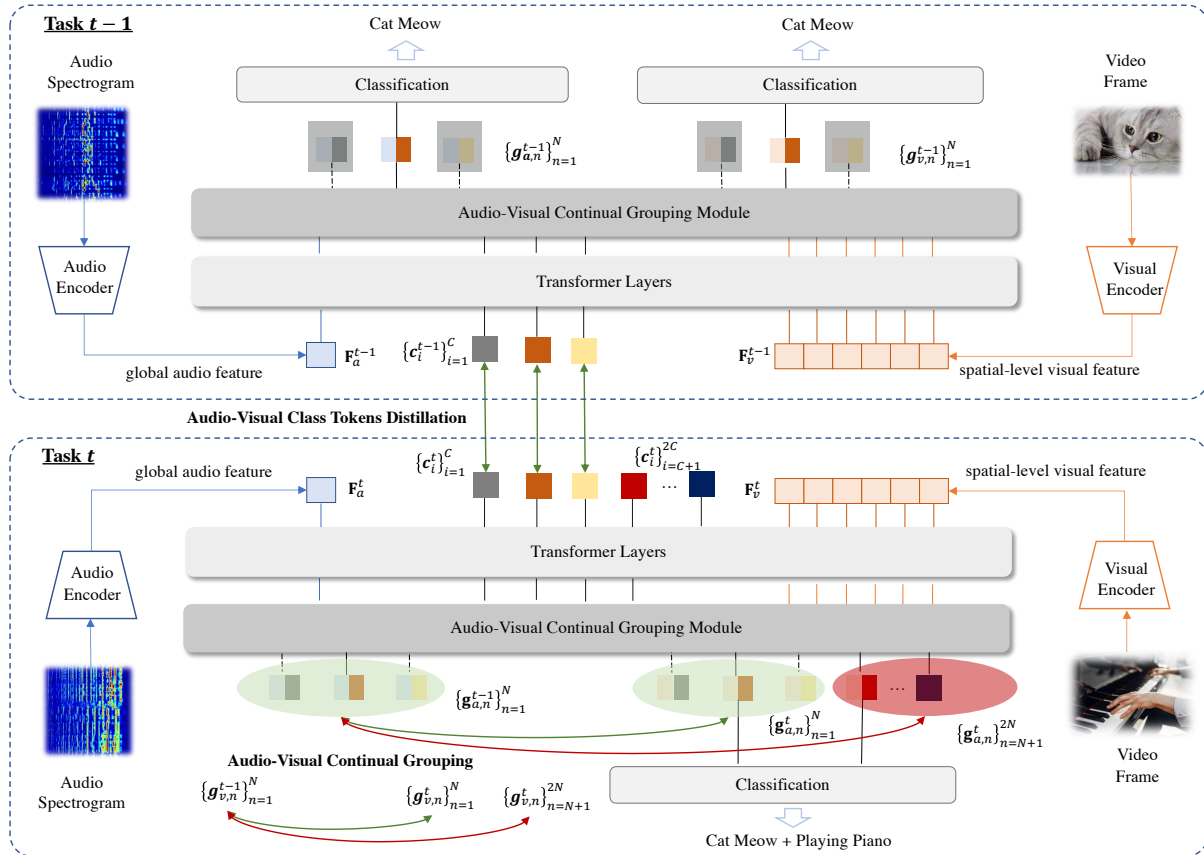
Figure 2. Illustration of the proposed Class-Incremental Grouping Network (CIGN). The Audio-Visual Continual Grouping module takes as audio features $\mathbf{F}_a^t = \mathbf{f}_a^t$ of the spectrogram, visual features $\mathbf{F}_v^t = \{\mathbf{f}_{v,p}^t\}_{p=1}^{P}$ of the video frame from each encoder and learnable audio-visual class tokens $\{\mathbf{c}_i^t\}_{i=1}^{2C}$ for $2C$ classes in the semantic space to generate disentangled class-aware audio-visual representations $\{\mathbf{g}_{a,n}^t\}_{n=1}^{2N}, \{\mathbf{g}_{v,n}^t\}_{n=1}^{2N}$ for $2N$ sources at task t. Note that $2N$ source embeddings are chosen from $2C$ categories according to the ground-truth class. The Audio-Visual Class Tokens Distillation module constrains the distribution of old class tokens $\{\mathbf{c}_i^t\}_{i=1}^{C}$ at task t same as those $\{\mathbf{c}_i^{t-1}\}_{i=1}^{C}$ at task $t-1$. Meanwhile, class-aware audio-visual features $\{\mathbf{g}_{a,n}^{t-1}\}_{n=1}^{N}, \{\mathbf{g}_{v,n}^{t-1}\}_{n=1}^{N}$ at task $t-1$ are pulled close (Green arrow) to the associated class-aware features $\{\mathbf{g}_{a,n}^t\}_{n=1}^{N}, \{\mathbf{g}_{v,n}^t\}_{n=1}^{N}$ and away (Red arrow) from those features with different classes $\{\mathbf{g}_{a,n}^t\}_{n=N+1}^{2N}, \{\mathbf{g}_{v,n}^t\}_{n=N+1}^{2N}$ at task t. Finally, the classification layer composed of an FC layer and a sigmoid function is separately used to predict audio and video classes, where the product of FC output logits is utilized to generate audio-visual categories.

lized as the desirable guidance for class-incremental audio-visual learning.

Due to the promising performance of prompts in transfer learning, recent works [63, 62] also have explored different frameworks to alleviate catastrophic forgetting in class-incremental visual learning. A prompt pool memory-based framework was designed in L2P [63] to prompt a pre-trained vision transformer [14] for learning sequential classification tasks dynamically. Inspired by the Complementary Learning Systems [36, 30] theory, DualPrompt [62] combined General-Prompt and Expert-Prompt for learning task-invariant and task-specific knowledge. While those prompts-based approaches achieve promising performance in continual visual learning, they can only handle the texts targeted for images, and audio-visual prompts need to be well-designed. Applying their prompting to our audio-

visual settings fails to learn compact class-aware cross-modal representations for audio-visual class-incremental learning. In contrast, we develop a fully novel framework to aggregate compact category-wise audio and visual source representations with explicit learnable source class tokens. To the best of our knowledge, we are the first to leverage an explicit grouping mechanism for continual audio-visual learning. Our experiments in Section 4.2 also validate the superiority of CIGN in all benchmarks for class-incremental audio-visual settings.

3. Method

Given an image and a spectrogram of audio, our target is to classify the visual sound source sequentially for continual audio-visual learning. We propose a novel class-

incremental grouping network, named CIGN, for disentangling individual semantics from the audio and image, which mainly consists of two modules, Audio-Visual Class Tokens Distillation in Section 3.2 and Audio-Visual Continual Grouping in Section 3.3.

3.1. Preliminaries

In this section, we first describe the problem setup and notations and then revisit the audio-visual class-incremental classification for continual learning.

Problem Setup and Notations. Given a spectrogram and an image, our goal is to continually classify non-stationary audio-visual pairs given sequential tasks for class-incremental learning. We have an audio-visual label for a video with C audio-visual event categories, denoted as $\{y_i\}_{i=1}^{C}$ with y_i for the ground-truth category entry i as 1. During training, we have video-level annotations as supervision. Therefore, we can leverage the video-level label for the audio mixture spectrogram and image to perform continual audio-visual learning.

Revisit Continual Visual Learning. Given a sequence of tasks $\mathcal{D} = \{\mathcal{D}_1, ..., \mathcal{D}_T\}$, where the t-th task $\mathcal{D}_t = \{(\mathbf{x}_i^t, y_i^t)\}_{i=1}^{n_t}$ contains tuples of the input sample $\mathbf{x}_i^t \in \mathcal{X}$ and its corresponding label $y_i^t \in \mathcal{Y}$. The goal of continual visual learning is to train a single model with parameters θ, *i.e.*, $f_\theta : \mathcal{X} \rightarrow \mathcal{Y}$, which enables it to predict the label $y = f_\theta(\mathbf{x}) \in \mathcal{Y}$ given an unseen test sample $\mathbf{x}$ from arbitrary tasks. Data from the previous tasks may not be seen anymore when training future tasks. Different from traditional continual visual learning, we try to involve audio-visual pairs samples together in each sequential task. Specifically, the t-th task $\mathcal{D}_t$ includes tuples of the input audio-visual pairs $\mathbf{a}_i^t \in \mathcal{A}, \mathbf{v}_i^t \in \mathcal{V}$ and its corresponding label $y_i^t \in \mathcal{Y}$. Our target is to train a single multi-modal model with parameters θ, *i.e.*, $g_\theta : \mathcal{A} \times \mathcal{V} \rightarrow \mathcal{Y}$, such that we can predict the audio-visual event category $y = g_\theta(\mathbf{a}, \mathbf{v}) \in \mathcal{Y}$ given an unseen test audio-visual pair $\mathbf{a}, \mathbf{v}$ from arbitrary tasks. When training future tasks, audio-visual pairs from the previous tasks might not be seen anymore.

However, such a multi-modal continual learning setting will pose the main challenge for class-incremental audio-visual classification. The global audio representation extracted from the sound is catastrophically forgotten by the cross-modal model, and thus they can not associate individual audio with the corresponding image in future audio-visual tasks. To tackle the challenge, we are inspired by [67] and propose a novel class-incremental grouping network, namely CIGN, for disentangling individual event semantics from the audio and image to achieve class-incremental audio-visual learning, as illustrated in Figure 2.

3.2. Audio-Visual Class Tokens Distillation

To explicitly learn disentangled semantics from incremental audio and images, we introduce a novel audio-visual class token distillation to constrain the distribution of previous class tokens $\{\mathbf{c}_i^t\}_{i=1}^{C}$ at task t same as those $\{\mathbf{c}_i^{t-1}\}_{i=1}^{C}$ at task $t - 1$. The learnable audio-visual incremental class tokens $\{\mathbf{c}_i^t\}_{i=1}^{2C}$ at task t further group semantic-aware information from audio-visual representations $\mathbf{f}_a^t, \{\mathbf{f}_{v,p}^t\}_{p=1}^{P}$, where $\mathbf{c}_i^t \in \mathbb{R}^{1 \times D}$, $2C$ is the total number of old (C) and new (C) classes at task t, and P denotes the number of total spatial locations in the feature map.

With the incremental categorical audio-visual tokens and raw representations, we first apply self-attention transformers $\phi_a^t(\cdot), \phi_v^t(\cdot)$ at task t to aggregate global audio and spatial visual representations from the raw input and align the cross-modal features with the class token embeddings as:

$$\hat{\mathbf{f}}_a^t, \{\hat{\mathbf{c}}_{a,i}^t\}_{i=1}^{2C} = \{\phi_a^t(\mathbf{x}_{a,j}^t, \mathbf{X}_a^t, \mathbf{X}_a^t)\}_{j=1}^{1+2C},$$
$$\mathbf{X}_a^t = \{\mathbf{x}_{a,j}^t\}_{j=1}^{1+2C} = [\mathbf{f}_a^t; \{\mathbf{c}_i^t\}_{i=1}^{2C}] \tag{1}$$

$$\{\hat{\mathbf{f}}_{v,p}^t\}_{p=1}^{P}, \{\hat{\mathbf{c}}_{v,i}^t\}_{i=1}^{2C} = \{\phi_v^t(\mathbf{x}_{v,j}^t, \mathbf{X}_v^t, \mathbf{X}_v^t)\}_{j=1}^{P+2C},$$
$$\mathbf{X}^v = \{\mathbf{x}_{v,j}^t\}_{j=1}^{P+2C} = [\{\mathbf{f}_{v,p}^t\}_{p=1}^{P}; \{\mathbf{c}_i^t\}_{i=1}^{2C}] \tag{2}$$

where $\hat{\mathbf{f}}^a, \hat{\mathbf{f}}_p^v, \hat{\mathbf{c}}_i^a, \hat{\mathbf{c}}_i^v \in \mathbb{R}^{1 \times D}$, and D denotes the dimension of embeddings. $[\,;\,]$ is the concatenation operator. The self-attention operator $\phi_a^t(\cdot)$ is formulated as:

$$\phi_a^t(\mathbf{x}_{a,j}^t, \mathbf{X}_a^t, \mathbf{X}_a^t) = \text{Softmax}(\frac{\mathbf{x}_{a,j}^t \mathbf{X}_a^{t\top}}{\sqrt{D}})\mathbf{X}_a^t \tag{3}$$

$$\phi_v^t(\mathbf{x}_{v,j}^t, \mathbf{X}_v^t, \mathbf{X}_v^t) = \text{Softmax}(\frac{\mathbf{x}_{v,j}^t \mathbf{X}_v^{t\top}}{\sqrt{D}})\mathbf{X}_v^t \tag{4}$$

Then, to avoid forgetting old class tokens $\{\mathbf{c}_i^t\}_{i=1}^{C}$ at task $t - 1$, we apply a Kullback-Leibler (KL) divergence loss $\text{KL}(\mathbf{c}_i^t \| \mathbf{c}_i^{t-1})$ on new task t and previous task $t - 1$. Meanwhile, to constrain the independence of each new class token $\{\mathbf{c}_i^t\}_{i=C+1}^{2C}$, we use a fully-connected (FC) layer and add a softmax operator to predict the new class probability: $\mathbf{e}_i^t = \text{Softmax}(\text{FC}(\mathbf{c}_i^t))$. Then, a cross-entropy loss $\sum_{i=1}^{C} \text{CE}(\mathbf{h}_i^t, \mathbf{e}_i^t)$ optimizes each new audio-visual category probability, where $\text{CE}(\cdot)$ denote cross-entropy loss; $\mathbf{h}_i^t$ is a one-hot encoding with its element as 1 in the target class entry i. Optimizing the KL and cross-entropy loss together will push the learned token embeddings discriminative.

3.3. Audio-Visual Continual Grouping

With the benefit of the above constraints on learning old and new categories, we propose a novel and explicit audio-visual continual grouping module composed of grouping

Method	Audio		Visual		Audio-Visual	
	Average Acc ↑(%)	Forgetting ↓(%)	Average Acc ↑(%)	Forgetting ↓(%)	Average Acc ↑(%)	Forgetting ↓(%)
EWC [28]	17.32	37.28	15.19	39.75	23.27	33.65
LwF [34]	26.37	32.56	25.21	33.16	32.16	27.13
iCaRL [53]	27.25	30.75	28.32	29.37	33.26	26.53
IL2M [6]	31.65	26.72	30.29	28.63	38.13	21.25
BiC [65]	38.23	22.38	35.72	24.86	45.23	16.38
PODNet [15]	42.27	19.52	40.51	21.09	47.17	14.87
SS-IL [2]	47.05	15.09	45.59	16.83	53.26	13.16
L2P [63]	53.06	11.05	51.57	11.39	59.15	9.76
DualPrompt [62]	57.27	8.28	56.35	8.52	63.32	6.03
CIGN (ours)	**62.53**	**5.57**	**60.37**	**5.98**	**70.06**	**2.62**
Upper-bound	65.23	–	63.57	–	73.68	–

Table 1. Quantitative results of audio, visual, and audio-visual continual learning on VGGSound-Instruments dataset.

blocks $g_a^t(\cdot), g_v^t(\cdot)$ to take the learned audio-visual incremental category tokens and aggregated features as inputs to generate category-aware incremental embeddings as:

$$
\begin{aligned}
\{\mathbf{g}_{a,i}^t\}_{i=1}^{2C} &= g_a^t(\{\hat{\mathbf{f}}_a^t, \{\hat{\mathbf{c}}_i^a\}_{i=1}^{2C}), \\
\{\mathbf{g}_{v,i}^t\}_{i=1}^{2C} &= g_v^t(\{\hat{\mathbf{f}}_{v,p}^t\}_{p=1}^P, \{\hat{\mathbf{c}}_{v,i}^t\}_{i=1}^{2C})
\end{aligned}
\tag{5}
$$

During the grouping stage, we merge all the audio-visual embeddings with the same category token into an updated class-aware audio-visual feature by computing the global audio similarity vector $\mathbf{A}_a^t \in \mathbb{R}^{1 \times 2C}$ and spatial visual similarity matrix $\mathbf{A}_v^t \in \mathbb{R}^{P \times 2C}$ between audio-visual features and $2C$ class tokens at task t via a softmax operation, which is formulated as

$$
\begin{aligned}
\mathbf{A}_{a,i}^t &= \text{Softmax}(W_{q,a}^t \hat{\mathbf{f}}_a^t \cdot W_{k,a}^t \hat{\mathbf{c}}_{a,i}^t), \\
\mathbf{A}_{v,p,i}^t &= \text{Softmax}(W_{q,v}^t \hat{\mathbf{f}}_{v,p}^t \cdot W_{k,v}^t \hat{\mathbf{c}}_{v,i}^t)
\end{aligned}
\tag{6}
$$

where $W_{q,a}^t, W_{k,a}^t \in \mathbb{R}^{D \times D}$ and $W_{q,v}^t, W_{k,v}^t \in \mathbb{R}^{D \times D}$ denote learnable weights of linear projections for audio-visual features and class tokens at task t, respectively. Using this global audio similarity vector and spatial visual similarity matrix, we calculate the weighted sum of audio-visual features belonged to generate the class-aware embeddings as:

$$
\begin{aligned}
\mathbf{g}_{a,i}^t &= g_a^t(\hat{\mathbf{f}}_a^t, \hat{\mathbf{c}}_{a,i}^t) = \hat{\mathbf{c}}_{a,i}^t + W_{o,a}^t \frac{\mathbf{A}_{a,i}^t W_{v,a}^t \hat{\mathbf{f}}_a^t}{\mathbf{A}_{a,i}^t}, \\
\mathbf{g}_{v,i}^t &= g_v^t(\{\hat{\mathbf{f}}_{v,p}^t\}_{p=1}^P, \hat{\mathbf{c}}_{v,i}^t) = \hat{\mathbf{c}}_{v,i}^t + W_{o,v}^t \frac{\sum_{p=1}^P \mathbf{A}_{v,p,i}^t W_{v,v}^t \hat{\mathbf{f}}_{p,v}^t}{\sum_{p=1}^P \mathbf{A}_{v,p,i}^t}
\end{aligned}
\tag{7}
$$

where $W_{o,a}^t, W_{v,a}^t \in \mathbb{R}^{D \times D}$ and $W_{o,v}^t, W_{v,v}^t \in \mathbb{R}^{D \times D}$ denote learned weights of linear projections for output and value in terms of audio-visual modalities at task t, separately. With class-aware audio-visual features $\{\mathbf{g}_{a,i}^t\}_{i=1}^{2C}, \{\mathbf{g}_{v,i}^t\}_{i=1}^{2C}$ as the inputs, we apply an FC layer and sigmoid operator on them to predict the binary probability: $p_{a,i}^t = \text{Sigmoid}(\text{FC}(\mathbf{g}_{a,i}^t)), p_{v,i}^t = \text{Sigmoid}(\text{FC}(\mathbf{g}_{v,i}^t))$ for ith class. By applying audio-visual incremental classes

$\{y_i^t\}_{i=1}^{2C}$ as the supervision and combining the constraint loss, we formulate a class-constrained grouping loss as:

$$
\begin{aligned}
\mathcal{L}_{\text{group}} = \sum_{i=1}^C \text{KL}(\mathbf{c}_i^t || \mathbf{c}_i^{t-1}) + \sum_{i=C+1}^{2C} \text{CE}(\mathbf{h}_i^t, \mathbf{e}_i^t) \\
+ \sum_{i=1}^{2C} \{\text{BCE}(y_i^t, p_{a,i}^t) + \text{BCE}(y_i^t, p_{v,i}^t)\}.
\end{aligned}
\tag{8}
$$

With the help of the introduced class-constrained objective, we generate category-aware audio-visual representations $\{\mathbf{g}_{a,i}^t\}_{i=1}^{2C}, \{\mathbf{g}_{v,i}^t\}_{i=1}^{2C}$ for audio-visual alignment. Note that global audio-visual features for $2N$ class tokens $\{\mathbf{g}_{a,n}^t\}_{n=1}^{2N}, \{\mathbf{g}_{v,n}^t\}_{n=1}^{2N}$ are chosen from $2C$ categories according to the associated ground-truth class. Therefore, the audio-visual similarities between old and new classes are computed by max-pooling audio-visual cosine similarities of class-aware audio-visual features $\{\mathbf{g}_{a,n}^{t-1}\}_{n=1}^N, \{\mathbf{g}_{v,n}^{t-1}\}_{n=1}^N$ at task $t-1$ and those features from old classes $\{\mathbf{g}_{a,n}^t\}_{n=1}^N, \{\mathbf{g}_{v,n}^t\}_{n=1}^N$ and new classes $\{\mathbf{g}_{a,n}^t\}_{n=N+1}^{2N}, \{\mathbf{g}_{v,n}^t\}_{n=N+1}^{2N}$ at task t. With these category-aware similarities, we formulate the continual audio-visual grouping loss as:

$$
\begin{aligned}
\mathcal{L}_{\text{ctl}} = &-\frac{1}{N} \sum_{n=1}^N \log \frac{\exp\left(\frac{1}{\tau}\texttt{sim}(\mathbf{g}_{a,n}^{t-1}, \mathbf{g}_{a,n}^t)\right)}{\sum_{m=N+1}^{2N} \exp\left(\frac{1}{\tau}\texttt{sim}(\mathbf{g}_{a,n}^{t-1}, \mathbf{g}_{a,m}^t)\right)} \\
&-\frac{1}{N} \sum_{n=1}^N \log \frac{\exp\left(\frac{1}{\tau}\texttt{sim}(\mathbf{g}_{v,n}^{t-1}, \mathbf{g}_{v,n}^t)\right)}{\sum_{m=N+1}^{2N} \exp\left(\frac{1}{\tau}\texttt{sim}(\mathbf{g}_{v,n}^{t-1}, \mathbf{g}_{v,m}^t)\right)}
\end{aligned}
\tag{9}
$$

The overall objective of our model is simply optimized in an end-to-end manner as:

$$
\mathcal{L} = \mathcal{L}_{\text{ctl}} + \mathcal{L}_{\text{group}}
\tag{10}
$$

During inference, we follow the prior work [63, 62] and use one single model with parameters trained at task t for evaluation, and the product of output logits $(p_{a,i}^t, p_{v,i}^t)$ from audio and visual modalities are utilized to predict the probability of audio-visual classes, that is $p_{av,i}^t = p_{a,i}^t \cdot p_{v,i}^t$.

Method	Audio		Visual		Audio-Visual	
	Average Acc ↑(%)	Forgetting ↓(%)	Average Acc ↑(%)	Forgetting ↓(%)	Average Acc ↑(%)	Forgetting ↓(%)
EWC [28]	15.17	43.05	18.53	41.25	21.52	39.16
LwF [34]	24.25	33.83	26.78	32.56	29.16	28.25
iCaRL [53]	23.55	35.32	27.31	33.07	28.37	31.02
IL2M [6]	26.32	27.56	30.25	25.83	32.65	23.19
BiC [65]	32.31	22.13	35.09	19.35	38.56	17.02
PODNet [15]	33.56	21.23	37.16	17.56	40.03	15.51
SS-IL [2]	36.21	18.75	40.57	16.82	43.16	13.89
L2P [63]	38.05	16.13	42.06	13.81	47.82	11.05
DualPrompt [62]	42.25	13.58	46.28	10.03	49.72	9.63
CIGN (ours)	**45.83**	**10.21**	**49.52**	**8.83**	**55.26**	**6.52**
Upper-bound	50.06	–	53.32	–	60.57	–

Table 2. Quantitative results of audio, visual, and audio-visual continual learning on VGGSound-100 dataset.

4. Experiments

4.1. Experimental Setup

Datasets. VGGSound-Instruments [26] includes 32k video clips of 10s lengths from 36 musical instrument classes, a subset of VGG-Sound [13], and each video only has one single instrument class annotation. Beyond this instrumental benchmark, we filter 97k video clips of 10s lengths from the original VGG-Sound [13], denoted as VGGSound-100, and consists of 100 categories, such as nature, animals, vehicles, human speech, dancing, playing instruments, etc. For large-scale incremental learning, we use the full VGG-Sound Source [12] with 150k video clips with 220 categories in the original VGG-Sound [13]. Each dataset is split into train/val/test sets with respective ratios of 0.8/0.1/0.1.

Evaluation Metrics. Following the prior work [2, 63, 62], we use the class-incremental setting of $T = 4$ sequential classification tasks with equal sizes of categories, where each task has a separate test set. With the common metrics in previous methods [63, 62], we apply Average accuracy and Forgetting for comprehensive evaluation. Higher Average accuracy is better, and lower Forgetting is better.

Implementation. For input frames, we resize the resolution to 224×224. For input audio, the log spectrograms are generated from $3s$ of audio with a sample rate of 22050Hz. Following the prior work [39, 38], we apply STFT to extract an input tensor of size 257×300 (257 frequency bands over 300 timesteps) by using 50ms windows at a hop size of 25ms. We follow previous work [24, 52, 12, 39, 38], we apply the lightweight ResNet18 [22] as the audio and visual encoder, and initialize them using weights pre-trained on 2M Flickr videos [5] using the state-of-the-art self-supervised source localization approach [39]. The input dimension D of embeddings is 512, and the total number P of patches for the 7×7 spatial map is 49. The depth of self-attention transformers $\phi^a(\cdot), \phi^v(\cdot)$ is 3 as default. We train the model for 100 epochs with a batch size of 128. The Adam optimizer [27] is used with a learning rate of $1e - 4$. Following the prior work [63], we randomly use 50 audio-

visual pairs per category from old tasks for the buffer size.

4.2. Comparison to Prior Work

In this work, we propose a novel and effective framework for continual audio-visual learning. In order to demonstrate the effectiveness of the proposed CIGN, we comprehensively compare it to previous continual learning baselines: 1) EWC [28] (2017'PNAS): a vanilla baseline that addressed catastrophic forgetting in neural networks; 2) LwF [34] (2018'TPAMI): a regularization-based approach for training new task data and preserving capabilities of old tasks; 3) iCaRL [53] (2017'CVPR): a class-incremental work that continuously used a sequential data stream for new classes; 4) IL2M [6] (2019'ICCV): a dual memory network that combined a bounded memory of the past images and a second memory for past class statistics; 5) BiC [65] (2019'CVPR): a bias correction method that tackled the imbalance between old and new classes; 6) POD-Net [15] (2020'ECCV): a spatial-based distillation framework with proxy vectors for each category; 7) SSIL [2] (2021'ICCV): a task-wise knowledge distillation network based on the separated softmax output layer; 8) L2P [63] (2022'CVPR): a strong prompts-based baseline optimizing the prompt pool memory as instructions for sequential tasks; 9) DualPrompt [62] (2022'ECCV): a recent strong baseline with General-Prompt and Expert-Prompt for task-invariant and task-specific knowledge; 10) Upper-bound: a usually supervised baseline training on the data of all tasks.

For the VGGSound-Instruments dataset, we report the quantitative comparison results in Table 1. As can be seen, we achieve the best results regarding all metrics for three class-incremental settings (audio, visual, and audio-visual) compared to previous continual learning approaches. In particular, the proposed CIGN superiorly outperforms DualPrompt [62], the current state-of-the-art continual learning baseline, by 5.26 Average Acc@Audio & 2.71 Forgetting@Audio, 4.02 Average Acc@Visual & 2.54 Forgetting@Visual, and 6.74 Average Acc@Audio-Visual & 3.41 Forgetting@Audio-Visual on three settings. Furthermore, we achieve significant performance gains compared to

Method	Audio		Visual		Audio-Visual	
	Average Acc ↑(%)	Forgetting ↓(%)	Average Acc ↑(%)	Forgetting ↓(%)	Average Acc ↑(%)	Forgetting ↓(%)
EWC [28]	9.28	59.03	13.51	52.16	16.36	46.72
LwF [34]	17.67	48.36	20.32	45.36	25.72	38.65
iCaRL [53]	18.56	46.28	21.03	44.52	23.56	42.35
IL2M [6]	23.07	41.76	26.51	36.32	30.25	31.86
BiC [65]	25.68	37.83	29.26	31.33	33.19	23.56
PODNet [15]	28.23	32.57	31.06	29.71	36.51	18.62
SS-IL [2]	31.56	28.16	35.52	26.62	40.21	15.23
L2P [63]	34.21	25.19	36.21	20.07	42.15	13.52
DualPrompt [62]	37.52	22.63	40.86	17.27	43.61	11.56
CIGN (ours)	**39.81**	**15.25**	**42.26**	**13.01**	**46.58**	**8.67**
Upper-bound	45.89	–	51.56	–	53.75	–

Table 3. Quantitative results of audio, visual, and audio-visual continual learning on VGG-Sound Sources dataset.

AVCTD	AVCG	Audio		Visual		Audio-Visual	
		Average Acc ↑(%)	Forgetting ↓(%)	Average Acc ↑(%)	Forgetting ↓(%)	Average Acc ↑(%)	Forgetting ↓(%)
✗	✗	36.72	19.23	39.35	17.23	45.05	15.26
✓	✗	42.59	13.21	47.12	10.72	51.02	9.75
✗	✓	40.16	15.03	43.56	13.69	48.72	11.57
✓	✓	**45.83**	**10.21**	**49.52**	**8.83**	**55.26**	**6.52**

Table 4. Ablation studies on Audio-Visual Class Tokens Distillation (AVCTD) and Audio-Visual Continual Grouping (AVCG).

L2P [63], the current state-of-the-art rehearsal-based baseline, which indicates the importance of extracting category-aware semantics from incremental audio-visual inputs as guidance for continual audio-visual learning. Meanwhile, the gap between our CIGN and the oracle performance of upper-bound using all data for training is the lowest compared to other baselines. These significant improvements demonstrate the superiority of our approach in audio-visual class-incremental learning.

In addition, significant gains in VGGSound-100 and VGG-Sound Sources benchmarks can be observed in Table 2 and Table 3. Compared to L2P [63], the current state-of-the-art rehearsal-based method, we achieve the results gains of 7.78 Average Acc@Audio, 7.46 Average Acc@Visual, and 7.44 Average Acc@Audio-Visual on VGGSound-100 dataset. Moreover, when evaluated on the challenging VGG-Sound Sources benchmark, the proposed method still outperforms L2P [63] by 5.60 Average Acc@Audio, 6.05 Average Acc@Visual, and 4.43 Average Acc@Audio-Visual. We also achieve highly better results against SS-IL [2], the task-wise knowledge distillation network based on separated softmax. These results demonstrate the effectiveness of our approach in learning disentangled semantics from incremental audio and images for continual audio-visual classification.

4.3. Experimental Analysis

In this section, we performed ablation studies to demonstrate the benefit of introducing the Audio-Visual Class Tokens Distillation and Audio-Visual Continual Grouping modules. We also conducted extensive experiments to explore a flexible number of incremental tasks, and learned disentangled category-aware audio-visual representations.

Audio-Visual Class Tokens Distillation & Audio-Visual Continual Grouping. In order to demonstrate the effectiveness of the introduced audio-visual class tokens distillation (AVCTD) and audio-visual continual grouping (AVCG), we ablate the necessity of each module and report the quantitative results on VGGSound-100 dataset in Table 4. As can be observed, adding AVCTD to the vanilla baseline highly increases the results of Average Acc (by 5.87@Audio, 7.77@Visual, and 5.97@Audio-Visual) and decreases the performance of forgetting (by 6.02@Audio, 6.51@Visual, and 5.51@Audio-Visual), which validates the benefit of category tokens distillation in extracting disentangled high-level semantics for class-incremental learning. Meanwhile, introducing only AVCG in the baseline increases the class-incremental learning performance regarding all metrics. More importantly, incorporating AVCTD and AVCG into the baseline significantly raises the performance of Average Acc by 9.11@Audio, 10.17@Visual, and 10.21@Audio-Visual, and reduces the results of Forgetting by 9.02@Audio, 8.40@Visual, and 8.74@Audio-Visual. These improving results validate the importance of audio-visual class tokens distillation and audio-visual continual grouping in extracting category-aware semantics from class-incremental audio and image for continual audio-visual learning.

Generalizing to Flexible Number of Incremental Tasks. In order to validate the generalizability of the proposed CIGN to a flexible number of incremental tasks, we transfer the model to continual training on 10 tasks with every 10 categories on the VGGSound-100 benchmark. We still
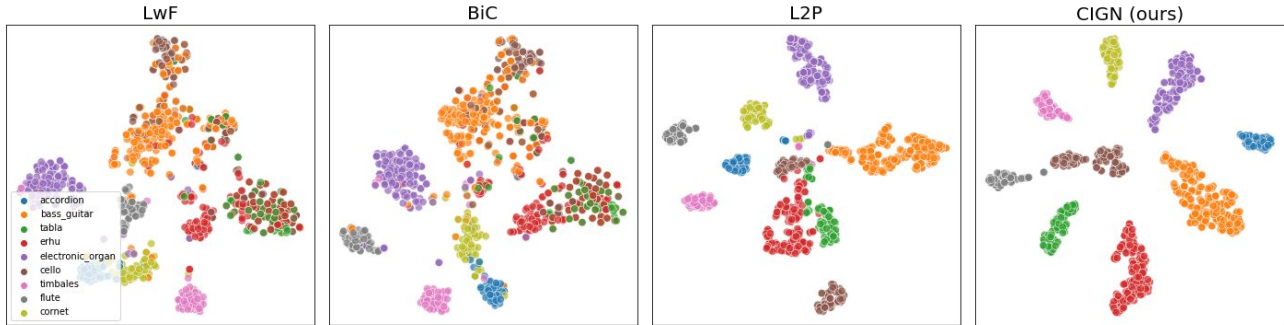
Figure 3. Qualitative comparisons of representations learned by LwF, BiC, L2P, and the proposed CIGN. Note that each spot denotes features extracted from one sample, and each color refers to one audio-visual category, such as "cello" in brown and "tabla" in green.

achieve competitive results of Average Acc (40.53@Audio, 43.29@Visual, 48.16@Audio-Visual) and Forgetting (13.87@Audio, 12.16@Visual, 9.75@Audio-Visual) on the challenging VGGSound-100 dataset. These results indicate that our CIGN can support a flexible number of incremental classification tasks, which further demonstrates the effectiveness of our class-incremental grouping network for continual audio-visual learning.

Learned Category-aware Audio-Visual Representations. Learning disentangled audio-visual representations with category-aware semantics is critical for classifying audio-visual pairs from incremental classes. To better evaluate the quality of learned category-aware features, we visualize the learned visual and audio representations of 9 categories in the first task after finishing 4 incremental tasks on VGGSound-Instruments by t-SNE [61], as shown in Figure 3. It should be noted that each color denotes one class of the audio-visual pair, such as "cello" in brown and "tabla" in green. As can be seen in the last column, audio-visual embeddings extracted by the proposed CIGN are both intra-class compact and inter-class separable. In contrast to our disentangled representations in the audio-visual semantic space, mixtures of multiple audio-visual categories still exist among features learned by LwF [34] and BiC [65]. With the help of text prompts for images, L2P [62] can extract clustered audio-visual incremental representations on some classes, such as "base_guitar" in orange. However, some categories are mixed while others are very close to each other as they do not incorporate the explicit audio-visual continual grouping mechanism in our CIGN. These meaningful visualization results further showcase the superiority of our CIGN in extracting compact audio-visual incremental representations with class-aware semantics for continual audio-visual learning.

4.4. Limitation

Although the proposed CIGN achieves superior results on both single-modality and cross-modal class-incremental learning, the performance gains of our method on the VGGSound-Instruments benchmark with a limited size of buffers are not substantial. One possible cause is that our model overfits during the training time, and the solution is to incorporate dropout and momentum encoders together for continual audio-visual classification. Meanwhile, we observe that if we transfer our model to open-set audio-visual classification without additional training, it would be hard to predict unseen categories as we need to pre-define a set of categories during training and do not learn unseen class tokens to guide the classification. Future work could add enough learnable category tokens for rehearsal-free continual learning in new categories.

5. Conclusion

In this work, we present CIGN, a novel class-incremental grouping network that can directly learn category-wise semantic features to achieve continual audio-visual learning. We leverage learnable audio-visual class tokens and audio-visual grouping to aggregate class-aware features continually. Furthermore, we introduce class tokens distillation and continual grouping to alleviate forgetting parameters learned from previous tasks for capturing discriminative audio-visual categories. Experimental results on VGGSound-Instruments, VGGSound-100, and VGG-Sound Sources benchmarks comprehensively demonstrate the state-of-the-art superiority against previous regularization- and rehearsal-based class-incremental learning baselines on continual audio-visual settings. Meanwhile, qualitative visualizations of incremental audio-visual embeddings vividly showcase the effectiveness of our CIGN in aggregating class-aware features to avoid cross-modal catastrophic forgetting. Extensive ablation studies also validate the importance of class tokens distillation and continual grouping in learning compact representations for continual audio-visual learning.

Broader Impact. The proposed method predicts class-incremental audio-visual pairs learning from web videos, which could cause the model to learn internal biases in the

data. For instance, the model might fail to predict certain rare but crucial audio-visual classes. These issues should be incrementally resolved for real-world applications.

References

[1] Triantafyllos Afouras, Andrew Owens, Joon Son Chung, and Andrew Zisserman. Self-supervised learning of audio-visual objects from video. In *Proceedings of European Conference on Computer Vision (ECCV)*, pages 208–224, 2020. 2

[2] Hongjoon Ahn, Jihwan Kwak, Subin Lim, Hyeonsu Bang, Hyojun Kim, and Taesup Moon. Ss-il: Separated softmax for incremental learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 844–853, 2021. 1, 2, 5, 6, 7

[3] Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 139–154, 2018. 2

[4] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 609–617, 2017. 2

[5] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2016. 2, 6

[6] Eden Belouadah and Adrian Popescu. Il2m: Class incremental learning with dual memory. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 583–592, 2019. 2, 5, 6, 7

[7] Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and SIMONE CALDERARA. Dark experience for general continual learning: a strong, simple baseline. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, pages 15920–15930, 2020. 2

[8] Francisco M. Castro, Manuel J. Marin-Jimenez, Nicolas Guil, Cordelia Schmid, and Karteek Alahari. End-to-end incremental learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 233–248, 2018. 1, 2

[9] Hyuntak Cha, Jaeho Lee, and Jinwoo Shin. Co2l: Contrastive continual learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9516–9525, 2021. 2

[10] Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet K Dokania, Philip HS Torr, and Marc'Aurelio Ranzato. On tiny episodic memories in continual learning. *arXiv preprint arXiv:1902.10486, 2019*, 2019. 2

[11] Changan Chen, Unnat Jain, Carl Schissler, S. V. A. Garí, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. Soundspaces: Audio-visual navigation in 3d environments. In *Proceedings of European Conference on Computer Vision (ECCV)*, pages 17–36, 2020. 2

[12] Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Localizing visual sounds the hard way. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16867–16876, 2021. 1, 2, 6

[13] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725. IEEE, 2020. 1, 6

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[15] Arthur Douillard, Matthieu Cord, Charles Ollion, Thomas Robert, and Eduardo Valle. Podnet: Pooled outputs distillation for small-tasks incremental learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, page 86–102, 2020. 2, 5, 6, 7

[16] Chuang Gan, Deng Huang, Hang Zhao, Joshua B. Tenenbaum, and Antonio Torralba. Music gesture for visual sound separation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10478–10487, 2020. 2

[17] Ruohan Gao, Rogerio Feris, and Kristen Grauman. Learning to separate object sounds by watching unlabeled video. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 35–53, 2018. 1, 2

[18] Ruohan Gao and Kristen Grauman. 2.5d visual sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 324–333, 2019. 2

[19] Ruohan Gao and Kristen Grauman. Co-separating sounds of visual objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3879–3888, 2019. 2

[20] Ruohan Gao and Kristen Grauman. Visualvoice: Audio-visual speech separation with cross-modal consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15495–15505, 2021. 2

[21] Ruohan Gao, Tae-Hyun Oh, Kristen Grauman, and Lorenzo Torresani. Listen to look: Action recognition by previewing audio. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10457–10467, 2020. 2

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 6

[23] Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. Learning a unified classifier incrementally via rebalancing. In *Proceedings of the IEEE/CVF Conference*

on Computer Vision and Pattern Recognition (CVPR), pages 831–839, 2019. 1, 2

[24] Di Hu, Feiping Nie, and Xuelong Li. Deep multimodal clustering for unsupervised audiovisual learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9248–9257, 2019. 2, 6

[25] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2022. 1

[26] Xixi Hu, Ziyang Chen, and Andrew Owens. Mix and localize: Localizing sound sources in mixtures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10483–10492, 2022. 1, 6

[27] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[28] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. 1, 2, 5, 6, 7

[29] Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 2

[30] Dharshan Kumaran, Demis Hassabis, and James L. McClelland. What learning systems do intelligent agents need? complementary learning systems theory updated. *Trends in Cognitive Sciences*, 20:512–534, 2016. 3

[31] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, 2021. 1

[32] Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. Learning to answer questions in dynamic audio-visual scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19108–19118, 2022. 2

[33] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021. 1

[34] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947, 2018. 1, 2, 5, 6, 7, 8

[35] Yan-Bo Lin, Hung-Yu Tseng, Hsin-Ying Lee, Yen-Yu Lin, and Ming-Hsuan Yang. Exploring cross-video and cross-modality signals for weakly-supervised audio-visual video parsing. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2

[36] James L. McClelland, Bruce L. McNaughton, and Randall C. O'Reilly. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological review*, 102(3):419–457, 1995. 3

[37] Michael McCloskey and Neal J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. *Psychology of Learning and Motivation*, 24:109–165, 1989. 1

[38] Shentong Mo and Pedro Morgado. A closer look at weakly-supervised audio-visual source localization. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2, 6

[39] Shentong Mo and Pedro Morgado. Localizing visual sounds the easy way. In *Proceedings of European Conference on Computer Vision (ECCV)*, page 218–234, 2022. 1, 2, 6

[40] Shentong Mo and Pedro Morgado. A unified audio-visual learning framework for localization, separation, and recognition. *arXiv preprint arXiv:2305.19458*, 2023. 2

[41] Shentong Mo, Jing Shi, and Yapeng Tian. DiffAVA: Personalized text-to-audio generation with visual alignment. *arXiv preprint arXiv:2305.12903*, 2023. 2

[42] Shentong Mo and Yapeng Tian. Multi-modal grouping network for weakly-supervised audio-visual video parsing. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2

[43] Shentong Mo and Yapeng Tian. Audio-visual grouping network for sound localization from mixtures. *arXiv preprint arXiv:2303.17056*, 2023. 2

[44] Shentong Mo and Yapeng Tian. AV-SAM: Segment anything model meets audio-visual localization and segmentation. *arXiv preprint arXiv:2305.01836*, 2023. 2

[45] Pedro Morgado, Yi Li, and Nuno Nvasconcelos. Learning representations from audio-visual spatial alignment. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, pages 4733–4744, 2020. 2

[46] Pedro Morgado, Ishan Misra, and Nuno Vasconcelos. Robust audio-visual instance discrimination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12934–12945, 2021. 2

[47] Pedro Morgado, Nuno Nvasconcelos, Timothy Langlois, and Oliver Wang. Self-supervised generation of spatial audio for 360°video. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 2

[48] Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. Audio-visual instance discrimination with cross-modal agreement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12475–12486, June 2021. 2

[49] Andrew Owens, Jiajun Wu, Josh H. McDermott, William T. Freeman, and Antonio Torralba. Ambient sound provides supervision for visual learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 801–816, 2016. 2

[50] Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. Adapterfusion: Non-destructive task composition for transfer learning. *arXiv preprint arXiv:2005.00247*, 2020. 1

[51] Ameya Prabhu, Philip Torr, and Puneet Dokania. Gdumb: A simple approach that questions our progress in continual learning. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2020. 1, 2

[52] Rui Qian, Di Hu, Heinrich Dinkel, Mengyue Wu, Ning Xu, and Weiyao Lin. Multiple sound sources localization from coarse to fine. In *Proceedings of European Conference on Computer Vision (ECCV)*, pages 292–308, 2020. 2, 6

[53] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5533–5542, 2017. 1, 2, 5, 6, 7

[54] Andrew Rouditchenko, Hang Zhao, Chuang Gan, Josh H. McDermott, and Antonio Torralba. Self-supervised audio-visual co-segmentation. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2357–2361, 2019. 2

[55] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4358–4366, 2018. 1, 2

[56] Kun Su, Xiulong Liu, and Eli Shlizerman. Audeo: Audio generation for a silent performance video. *Advances in Neural Information Processing Systems*, 33:3325–3337, 2020. 2

[57] Kun Su, Xiulong Liu, and Eli Shlizerman. How does it sound? *Advances in Neural Information Processing Systems*, 34:29258–29273, 2021. 2

[58] Yapeng Tian, Di Hu, and Chenliang Xu. Cyclic co-learning of sounding object visual grounding and sound separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2745–2754, 2021. 1, 2

[59] Yapeng Tian, Dingzeyu Li, and Chenliang Xu. Unified multisensory perception: Weakly-supervised audio-visual video parsing. In *Proceedings of European Conference on Computer Vision (ECCV)*, page 436–454, 2020. 2

[60] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2018. 1, 2

[61] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. 8

[62] Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, et al. Dualprompt: Complementary prompting for rehearsal-free continual learning. In *Proceedings of European Conference on Computer Vision (ECCV)*, page 631–648, 2022. 2, 3, 5, 6, 7, 8

[63] Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 139–149, 2022. 2, 3, 5, 6, 7

[64] Yake Wei, Di Hu, Yapeng Tian, and Xuelong Li. Learning in audio-visual context: A review, analysis, and new perspective. *arXiv preprint arXiv:2208.09579*, 2022. 2

[65] Yue Wu, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, and Yun Fu. Large scale incremental learning. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 374–382, 2019. 1, 2, 5, 6, 7, 8

[66] Yu Wu and Yi Yang. Exploring heterogeneous clues for weakly-supervised audio-visual video parsing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1326–1335, 2021. 2

[67] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. *arXiv preprint arXiv:2202.11094*, 2022. 4

[68] Hang Zhao, Chuang Gan, Wei-Chiu Ma, and Antonio Torralba. The sound of motions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1735–1744, 2019. 2

[69] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 570–586, 2018. 1, 2