Audio-Visual Class-Incremental Learning

Weiguo Pian[1†], Shentong Mo[2†], Yunhui Guo[1], Yapeng Tian[1]
[1] The University of Texas at Dallas, [2] Carnegie Mellon University
{weiguo.pian,yunhui.guo,yapeng.tian}@utdallas.edu, shentonm@andrew.cmu.edu

Abstract

In this paper, we introduce audio-visual class-incremental learning, a class-incremental learning scenario for audio-visual video recognition. We demonstrate that joint audio-visual modeling can improve class-incremental learning, but current methods fail to preserve semantic similarity between audio and visual features as incremental step grows. Furthermore, we observe that audio-visual correlations learned in previous tasks can be forgotten as incremental steps progress, leading to poor performance. To overcome these challenges, we propose AV-CIL, which incorporates Dual-Audio-Visual Similarity Constraint (D-AVSC) to maintain both instance-aware and class-aware semantic similarity between audio-visual modalities and Visual Attention Distillation (VAD) to retain previously learned audio-guided visual attentive ability. We create three audio-visual class-incremental datasets, AVE-Class-Incremental (AVE-CI), Kinetics-Sounds-Class-Incremental (K-S-CI), and VGGSound100-Class-Incremental (VS100-CI) based on the AVE, Kinetics-Sounds, and VGGSound datasets, respectively. Our experiments on AVE-CI, K-S-CI, and VS100-CI demonstrate that AV-CIL significantly outperforms existing class-incremental learning methods in audio-visual class-incremental learning. Code and data are available at: https://github.com/weiguoPian/AV-CIL_ICCV2023.

1. Introduction

Human perception of the environment is based on a variety of senses. Specifically, people perceive the world by seeing and listening to the events happening around them, which are two of the most commonly used signals for environment perception [5, 64, 63]. By jointly receiving visual and auditory signals, the human brain can better understand its surroundings. Researchers have been inspired by this practical approach to real-world perception and have begun to focus on audio-visual scene understanding. The goal is to guide machines to better perceive their surroundings by learning from audio and visual information jointly, just as humans do. In recent years, a lot of works have been explored in the field of audio-visual scene understanding, such as audio-visual event localization [51, 82, 83, 91], audio-visual video parsing [52, 64, 81, 90], audio-visual sound separation [25, 26, 80, 98], audio-visual video captioning [73, 79, 87], and audio-visual sound source localization [15, 18, 34, 35, 47, 61, 76]. These works have shown that modeling audio-visual modalities jointly can capture cross-modal semantic correlations effectively. Motivated by the success of these works, we aim to utilize the advantages of audio-visual modeling to mitigate the catastrophic forgetting problem [50, 57, 74] in class-incremental learning.

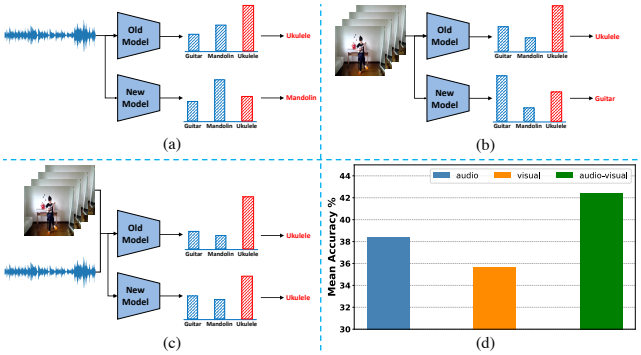


Figure 1: Illustration of training the model in a class-incremental manner using (a) audio modality, (b) visual modality, and (c) joint audio-visual modalities. (d) Mean accuracy of using audio modality, visual modality, and joint audio-visual modalities in class-incremental learning with a vanilla fine-tuning strategy on the AVE-CI dataset. Our results show that joint audio-visual modeling can significantly improve perception in the class-incremental setting.

Catastrophic forgetting means the model's performance on previous classes/tasks degrades significantly when updating it by training with only new classes'/tasks' data (or combined with a small set of previous classes'/tasks' data), which is the key challenge in incremental/continual learning. To alleviate it, a lot of works have been conducted

[†]Equal contribution.

in recent years, which can be categorized into parameter regularization-based [1, 4, 9, 40, 45, 96], knowledge distillation-based [6, 19, 22, 33, 41, 46, 50, 74, 89, 97, 99, 100], replay-memory/exemplar-based [2, 10, 11, 41, 53, 74, 77], and dynamic architecture-based methods [27, 37, 49, 86, 93, 94], mainly focus on continual/incremental image classification [2, 22, 74, 99], action recognition [48, 70], semantic segmentation [7, 8, 21, 29, 55, 68], object detection [20, 23, 39], language/joint-vision-language tasks [43, 59, 75, 78], and self-supervised representation learning/pre-training [24, 38, 54, 72, 92]. Despite the success of audio-visual modeling in capturing cross-modal semantic correlations, its potential in addressing the catastrophic forgetting problem in class-incremental learning remains unexplored.

In order to explore the effectiveness of joint audio-visual modeling in class-incremental learning, we first train the model using a vanilla fine-tuning strategy, which involves fine-tuning the model on new classes directly without any specific techniques (please see Section 4 for experimental settings). The results are presented in Figure 1, which demonstrates the advantage of joint audio-visual modeling over single audio or visual modality in a class-incremental manner. Based on these findings, we propose Audio-Visual Class-Incremental Learning, which is a novel incremental learning problem under the scenario of audio-visual video recognition. Since both audio and visual modalities are involved, simply applying existing class-incremental learning approaches to our new audio-visual task cannot fully exploit the natural cross-modal association between the two modalities. Instead, we propose that the correlation preservation between visual and audio modalities should be explicitly incorporated to further improve the current techniques for incremental learning of audio-visual data.

In this paper, we propose a method named AV-CIL (**A**udio-**V**isual-**C**lass-**I**ncremental **L**earning) to address our new audio-visual class-incremental learning problem. In AV-CIL, we introduce a *Dual-Audio-Visual Similarity Constraint (D-AVSC)*, which is designed to preserve both instance-aware and class-aware semantic similarities between audio and visual modalities throughout the increase in the number of classes.

Moreover, to better learn the cross-modal feature correlations, and ultimately get better joint audio-visual features, we also adopt an audio-guided visual attention mechanism [47, 82, 84, 91], which is an effective attention mechanism for adaptively learning correlations between audio and visual features. However, under the incremental learning settings, with the increasing of the incremental steps, we observe that the learned audio-visual attentive ability in previous tasks could get vanished, which results in the forgetting of learned audio-visual correlations in previous tasks. Please see Figure 3 for the visualization of this phenomenon. To preserve and leverage the previously learned attentive ability in future classes/tasks, we propose the *Visual Attention Distillation (VAD)* to distil the learned audio-guided visual attentive ability into new incremental steps, which enables the model to preserve previously learned cross-modal audio-visual correlations in new classes/tasks.

We use three existing audio-visual datasets: AVE [82], Kinetics-Sounds [3], and VGGSound [16] to construct datasets for class-incremental learning. We name the three newly constructed datasets as AVE-Class-Incremental (AVE-CI), Kinetics-Sounds-Class-Incremental (K-S-CI), and VGGSound100-Class-Incremental (VS100-CI), respectively. We conduct experiments on the three datasets to evaluate the effectiveness of our method in audio-visual class-incremental learning. The experimental results show that our proposed method outperforms state-of-the-art class-incremental learning methods significantly on all three datasets. In summary, this paper contributes follows:

- To explore the effectiveness of joint audio-visual modeling in the alleviation of the catastrophic forgetting problem in class-incremental learning, we propose audio-visual class-incremental learning that trains the model continually under the scenario of audio-visual video recognition. To the best of our knowledge, this is the first work on audio-visual incremental learning.

- We propose a method, named *AV-CIL*, to tackle the posed new problem. AV-CIL contains a *Dual-Audio-Visual Similarity Constraint (D-AVSC)* to preserve both instance- and class-aware semantic similarity between audio and visual features throughout the incremental steps. Furthermore, we also propose *Visual Attention Distillation (VAD)* to enable the model to preserve previously learned attentive ability in future classes/tasks for preventing the model from forgetting previously learned audio-visual correlations.

- Experimental results on three audio-visual class-incremental datasets, AVE-CI, K-S-CI, and VS100-CI (constructed from AVE [82], Kinetics-Sounds [3], and VGGSound [16]), demonstrate that our method outperforms state-of-the-art class-incremental learning methods significantly.

2. Related Work

2.1. Audio-Visual Learning

Audio-visual data can provide more synchronized and/or complementary information than unimodal audio or visual data, which has been proven to be more effective in scene understanding tasks, such as visual-guided sound separation/localization [15, 18, 25, 26, 34, 35, 47, 60, 61, 62, 65, 66, 76, 80, 98], audio-visual video parsing [52, 64, 81, 90], audio-visual event localization [51, 82, 83, 91], audio-

visual video caption [73, 79, 87], audio-visual navigation [12, 13, 14], etc. Lately, researchers have also been paying attention to generalizable audio-visual representation learning methods, *e.g.* audio-visual zero-shot learning methods [56, 58], and joint audio-visual pre-training [28]. A recent survey on audio-visual learning can be found in [88]. In this work, we mainly focus on audio-visual learning under class-incremental settings, in which the model continually learns data from new tasks/classes, to explore the feasibility of using cross-modal audio-visual correlations to mitigate the catastrophic forgetting problem in class-incremental learning.

2.2. Incremental Learning

Parameter Regularization. Parameter regularization-based methods [1, 4, 9, 40, 45, 96] aims to estimate the importance of different parameters of the model and allocate them with different weights to indicate their importance. During incremental steps, unimportant parameters can be updated much easier than important parameters.

Knowledge Distillation. Knowledge distillation-based methods [6, 19, 22, 33, 41, 46, 50, 74, 89, 97, 99, 100] help the model preserve previously learned knowledge in current/future incremental steps, which can be implemented as the minimization of the distance between the representations generated by the previous and current model [22, 41], or by minimizing the divergence (*e.g.* Kullback-Leibler divergence) between the output probability distribution of the previous and current model [2, 50, 74].

Exemplar/Memory Replay. Replay-based methods assume that small size of memory is accessible to store examples from old tasks/classes [2, 10, 11, 41, 53, 74, 77]. iCaRL [74] first proposed the nearest-mean-of-exemplars selection strategy to select the most representative exemplars in each class, which is also followed by later works [41, 70]. On the other hand, pseudo-rehearsal techniques [67, 69] were also proposed to use generative models to generate pseudo-exemplars based on the estimated distribution of data from previous classes.

Dynamic Architecture. Architecture-based methods [27, 37, 49, 86, 93, 94] hold incremental modules to increase the capacity of the model to handle new tasks/classes. Conventional architecture-based methods may cause unaffordable overhead caused by adding new modules continually as the incremental step grows [93]. Recently, the combination between dynamic architecture and distillation alleviates the continual-increasing overhead problem by distilling the increased modules into the original volume [86].

Existing Fields of Incremental Learning. The aforementioned incremental learning methods or techniques mainly focus on the fields of image classification [2, 22, 30, 74, 95, 99], action recognition [48, 70], semantic segmentation [7, 8, 21, 29, 55, 68], object detection [20, 23, 39], language/joint-vision-language tasks [43, 59, 75, 78], and self-supervised representation learning/pre-training [24, 38, 54, 72, 92]. In this paper, we focus on audio-visual class-incremental learning, in which we try to use the strengths of joint audio-visual modeling to alleviate the catastrophic forgetting problem in class-incremental learning.

3. Method

3.1. Problem Formulation

Class-incremental learning aims to train the model $\mathcal{F}_{\mathbf{\Theta}}$ with parameters $\mathbf{\Theta}$ through a sequence of T tasks $\{\mathcal{T}_1, \mathcal{T}_2, ..., \mathcal{T}_T\}$. In our audio-visual class-incremental learning, for a task $\mathcal{T}_t$ (incremental step t), its corresponding training set can be denoted as $\mathcal{D}_t = \{(\boldsymbol{x}^a_{t,i}, \boldsymbol{x}^v_{t,i}, y_{t,i})\}_{i=1}^{n_t}$, where $\boldsymbol{x}^a_{t,i}$ and $\boldsymbol{x}^v_{t,i}$ denote the i^{th} input sample's audio and visual modalities respectively in $\mathcal{D}_t$, and $y_{t,i} \in \mathcal{C}_t$ is the corresponding label, where $\mathcal{C}_t$ denotes the label space of task $\mathcal{T}_t$. For any two tasks' training label space $\mathcal{C}_{t_1}$ and $\mathcal{C}_{t_2}$, we have $\mathcal{C}_{t_1} \cap \mathcal{C}_{t_2} = \emptyset$. In our settings, storing a small fixed size of exemplar/memory set of data from previous classes is permitted, which is denoted as $\mathcal{M}_t$. Therefore, all accessible training data at incremental step t can be denoted as $\mathcal{D}'_t = \mathcal{D}_t \cup \mathcal{M}_t$. Note that, $|\mathcal{M}_1| = 0$, and $|\mathcal{M}_t|$ is a fixed number when $t > 1$. We use $|\mathcal{C}_t|$ and $|\mathcal{C}_{\mathcal{M}_t}|$ to denote the number of classes in task $\mathcal{T}_t$ and the number of classes in task $\mathcal{T}_t$'s exemplar set (the number of classes in task $\mathcal{T}_t$'s all previous tasks) respectively. Therefore, the exemplar number of each class in $\mathcal{M}_t$ can be presented as $m_t = |\mathcal{M}_t|/|\mathcal{C}_{\mathcal{M}_t}|$. We also let $|\mathcal{C}'_t|$ be the number of all classes up to the incremental step t, *i.e.* $|\mathcal{C}'_t| = |\mathcal{C}_t| + |\mathcal{C}_{\mathcal{M}_t}|$. Thus, the training process at incremental step t can be formulated as:

$$\mathbf{\Theta}_t = \underset{\mathbf{\Theta}_{t-1}}{\operatorname{argmin}} \mathbb{E}_{(\boldsymbol{x}^a, \boldsymbol{x}^v, y) \sim \mathcal{D}'_t} [\mathcal{L}(\mathcal{F}_{\mathbf{\Theta}_{t-1}}(\boldsymbol{x}^a, \boldsymbol{x}^v), y)], \quad (1)$$

where $\mathcal{L}$ is the loss function between the model's outputs and the corresponding labels. After the training process on $\mathcal{D}'_t$, the model is evaluated on the testing set that contains data from all seen classes up to step t.

3.2. Overview

The overview of our proposed method, AV-CIL, is illustrated in Figure 2. Our method mainly consists of Task-wise Knowledge Distillation [2, 50], Separated Softmax Cross-Entropy [2], and our proposed Dual-Audio-Visual Similarity Constraint (D-AVSC) and Visual Attention Distillation (VAD). The model used in our method is composed of a visual encoder, an audio encoder, an audio-guided visual attention layer [47, 82, 84, 91], an audio-visual fusion layer, and a classifier. For the audio encoder and the visual encoder, inspired by the excellent performance and generalization capability of recent vision/audition self-supervised pre-trained models [17, 31, 32, 36, 85], we apply

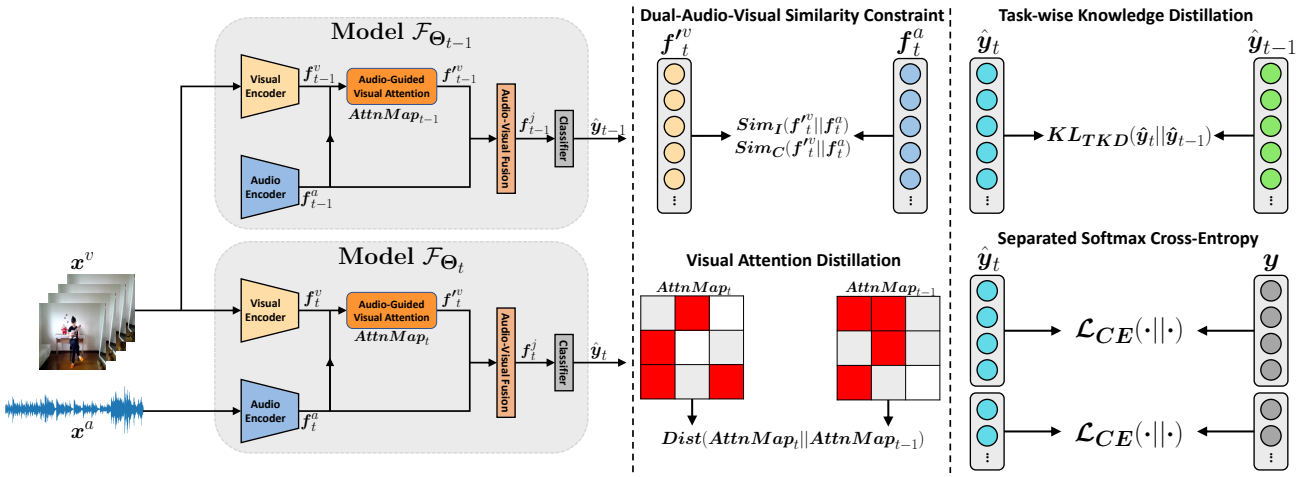


Figure 2: Overview of our proposed AV-CIL, which consists of four main components: Dual-Audio-Visual Similarity Constraint (D-AVSC), Visual Attention Distillation (VAD), Task-wise Knowledge Distillation [2, 50], and Separated Softmax Cross-Entropy [2].

the self-supervised pre-trained AudioMAE [36] and VideoMAE [85] as the audio encoder and the visual encoder, respectively. Moreover, due to their label- and class-free pre-training scheme, the self-supervised pre-trained models are suitable for incremental learning settings.

Given an input sequence of frames $\boldsymbol{x}^v$ and the associated audio signal $\boldsymbol{x}^a$, we first use the audio encoder (AudioMAE) and the visual encoder (VideoMAE) to generate the audio feature $\boldsymbol{f}^a \in \mathbb{R}^d$ and the visual feature $\boldsymbol{f}^v \in \mathbb{R}^{L\times S\times d}$ respectively, where L and S denote the temporal and spatial dimension of the visual feature respectively. After that, the audio-guided visual attention layer is used to pool the visual feature by both spatial and temporal dimensions, which is an effective mechanism to *adaptively* learn correlations between audio and visual features [47, 82, 84, 91].

Firstly, we calculate the spatial attention map on each temporal frame through the following process:

$$\begin{aligned} \boldsymbol{Score}^a &= \sigma(\boldsymbol{f}^a\boldsymbol{W}^a), \\ \boldsymbol{Score}^v_l &= \sigma(\boldsymbol{f}^v_l\boldsymbol{W}^v), \\ \boldsymbol{w}^{Spa.}_l &= Softmax(\boldsymbol{Score}^a \odot \boldsymbol{Score}^v_l), \end{aligned} \tag{2}$$

where $\boldsymbol{W}^a \in \mathbb{R}^{d\times d}$ and $\boldsymbol{W}^v \in \mathbb{R}^{d\times d}$ are the learnable projection matrices for audio and visual features respectively, and $\sigma(\cdot)$ is the nonlinear activation unit, *e.g.* Tanh function. $\boldsymbol{f}^v_l \in \mathbb{R}^{S\times d}$ and $\boldsymbol{w}^{Spa.}_l \in \mathbb{R}^{S\times d}$ denote the l^{th} temporal frame's feature map and spatial attention map respectively, $\odot$ denotes the Hadamard product. Based on the calculated spatial attention map of each temporal frame, the temporal attention map calculation can be formulated as:

$$\begin{aligned} \boldsymbol{Score'}^v_l &= \sum(\boldsymbol{w}^{Spa.}_l \odot \boldsymbol{Score}^v_l), \\ \boldsymbol{w}^{Tem.} &= Softmax([\boldsymbol{Score'}^v_1, ..., \boldsymbol{Score'}^v_L]), \end{aligned} \tag{3}$$

where $\boldsymbol{Score'}^v_l \in \mathbb{R}^d$. $\boldsymbol{w}^{Tem.} \in \mathbb{R}^{L\times d}$ denotes the temporal attention map. With the calculated spatial and temporal attention map, we can pool the original visual feature to get the attend visual feature map with more audio-visual cross-modal correlations:

$$\boldsymbol{f'}^v = \sum_{l=1}^{L} \boldsymbol{w}^{Tem.}_l \odot \sum(\boldsymbol{f}^v_l \odot \boldsymbol{w}^{Spa.}_l), \tag{4}$$

After the above process, an audio-visual fusion layer follows to get the joint audio-visual features that will be fed into the classifier to get the final results:

$$\hat{y} = CLS(\sigma(\boldsymbol{f}^a\boldsymbol{U}^a) + \sigma(\boldsymbol{f'}^v\boldsymbol{U}^v)), \tag{5}$$

where $\boldsymbol{U}^a \in \mathbb{R}^{d\times d}$ and $\boldsymbol{U}^v \in \mathbb{R}^{d\times d}$ are the projection matrices, and CLS denotes the classifier.

3.3. Dual-Audio-Visual Similarity Constraint

In this subsection, we will introduce our proposed Dual-Audio-Visual Similarity Constraint (D-AVSC), which aims to preserve the cross-modal semantic similarity, which is crucial to audio-visual modeling [64, 84], throughout the incremental steps.

Our D-AVSC is a dual constraint consisting of an Instance-aware Audio-Visual Semantic Similarity (I-AVSS) and a Class-aware Audio-Visual Semantic Similarity (C-AVSS). In I-AVSS, our goal is to maximize the similarity between audio and visual semantic features extracted from the *same video sample*, while minimizing the cross-modal similarity between audio and visual semantic features obtained from *different video samples*. In the incremental step

t, where $t > 1$, our I-AVSS can be formulated as:

$$\mathcal{L}_I = -\mathbb{E}_{(\boldsymbol{x}_i^a, \boldsymbol{x}_i^v)\sim\mathcal{D}_t'}\left[\log\frac{e^{\boldsymbol{f}_{t,i}^a \boldsymbol{f'}_{t,i}^{v\,\mathrm{T}}/\tau}}{\sum_{j=1}^{N} e^{\boldsymbol{f}_{t,i}^a \boldsymbol{f'}_{t,j}^{v\,\mathrm{T}}/\tau}}\right], \quad (6)$$

where N denotes the number of samples in a mini-batch. Our C-AVSS aims to preserve the cross-modal semantic similarities/dissimilarities among samples from the same/different classes. Specifically, it can maximize the similarity between audio and visual semantic features belonging to the *same class* and minimize the similarity between audio and visual features from *different classes*. Our C-AVSS can be expressed as follows during the incremental step t:

$$\mathcal{L}_C = -\mathbb{E}_{(\boldsymbol{x}_i^a, \boldsymbol{x}_i^v, y_i)\sim\mathcal{D}_t'} \left[\log\frac{\sum_{j=1}^{N} e^{\boldsymbol{f}_{t,i}^a \boldsymbol{f'}_{t,i}^{v\,\mathrm{T}}/\tau}\cdot \mathbb{1}[y_i = y_j]}{(\sum_{j=1}^{N}\mathbb{1}[y_i = y_j])(\sum_{j=1}^{N} e^{\boldsymbol{f}_{t,i}^a \boldsymbol{f'}_{t,i}^{v\,\mathrm{T}}/\tau})}\right], \quad (7)$$

where $\mathbb{1}[y_i = y_j]$ is an indicator that equals to 1 iff $y_i = y_j$. Our Dual-Audio-Visual Similarity Constraint (D-AVSC) is formulated as:

$$\mathcal{L}_{D-AVSC} = \lambda_I \mathcal{L}_I + \lambda_C \mathcal{L}_C, \quad (8)$$

where λ_I and λ_C are hyperparameters to balance the two constraint loss values.

3.4. Visual Attention Distillation

As we mentioned before, the learned audio-visual correlation could vanish as the incremental step increases. We visualize this phenomenon in Figure 3, which is generated by the model trained without our proposed Visual Attention Distillation process. The example is randomly selected from the validation set. In the figure, the heatmap in each frame shows the attentive level of different spatial regions by the corresponding audio, and the transparency of each frame demonstrates its temporal attentive level. We can see a man playing guitar in the video. Before classes increase, the most attentive region in each frame is around the area where the hand strums strings. However, as the incremental step grows, the most attentive region in each frame changes, as well as the temporal attentive level of each temporal frame. To address this problem, we propose the *Visual Attention Distillation (VAD)* to enable the model to preserve and leverage the previously learned attentive ability in future classes/tasks. In this way, the model can still maintain the visual attention ability without forgetting. For the full version of the visualization, please see Appendix for details.

Specifically, during an incremental step t where $t > 1$, we use the spatial and temporal visual attention maps generated by the model learned from the last incremental step $t-1$ as the target for the knowledge distillation process of current learned attention maps. Note that we only distill the visual attention map for data from the exemplar set. Our VAD is composed of the spatial distillation part and the temporal distillation part. For the spatial distillation part, we use the following formulation to express it:

$$Dist_{spa.} = \mathbb{E}_{(\boldsymbol{x}^a, \boldsymbol{x}^v)\sim\mathcal{M}_t}\Big[KL(\boldsymbol{w}_t^{Spa.}||\boldsymbol{w}_{t-1}^{Spa.})\Big], \quad (9)$$

where $\boldsymbol{w}_t^{Spa.} = [\boldsymbol{w}_{t,1}^{Spa.}, \boldsymbol{w}_{t,2}^{Spa.}, ..., \boldsymbol{w}_{t,L}^{Spa.}]$ and $\boldsymbol{w}_{t-1}^{Spa.} = [\boldsymbol{w}_{t-1,1}^{Spa.}, \boldsymbol{w}_{t-1,2}^{Spa.}, ..., \boldsymbol{w}_{t-1,L}^{Spa.}]$ denote the spatial attention maps for all L temporal frames generated by model $\mathcal{F}_{\boldsymbol{\Theta}_t}$ and model $\mathcal{F}_{\boldsymbol{\Theta}_{t-1}}$ respectively. $KL(\cdot||\cdot)$ denotes the Kullback-Leibler (KL) divergence function. Similarly, the temporal distillation part can be formulated as:

$$Dist_{tem.} = \mathbb{E}_{(\boldsymbol{x}^a, \boldsymbol{x}^v)\sim\mathcal{M}_t}\Big[KL(\boldsymbol{w}_t^{Tem.}||\boldsymbol{w}_{t-1}^{Tem.})\Big], \quad (10)$$

where $\boldsymbol{w}_t^{Tem.}$ and $\boldsymbol{w}_{t-1}^{Tem.}$ are the temporal attention maps generated by model $\mathcal{F}_{\boldsymbol{\Theta}_t}$ and model $\mathcal{F}_{\boldsymbol{\Theta}_{t-1}}$ respectively. Finally, the overall of our VAD can be illustrated by combining the above two parts:

$$\mathcal{L}_{VAD} = \lambda_{VAD} Dist_{spa.} + (1 - \lambda_{VAD}) Dist_{tem.}. \quad (11)$$

3.5. Final Loss Function

Above, we introduce our proposed Dual-Audio-Visual Similarity Constraint (D-AVSC) and Visual Attention Distillation (VAD). For our final optimization objective, we combine our D-AVSC and VAD with the Separated Softmax Cross-Entropy (SS-CE) [2] and the Task-wise Knowledge Distillation (TKD) [2, 6, 50].

SS-CE [2] is a modified version of the original softmax loss function for class-incremental learning, aims to prevent the prediction score of old class from being penalized by training on new tasks, which can be denoted as:

$$\mathcal{L}_{SS-CE} = \mathbb{E}_{(\boldsymbol{x}^a, \boldsymbol{x}^v, y)\sim\mathcal{D}_t'}\Big[\mathcal{L}_{CE}(\hat{\boldsymbol{y}}^{\mathcal{C}_t}, \boldsymbol{y}^{\mathcal{C}_t})\cdot\mathbb{1}[y\in\mathcal{C}_t] + \mathcal{L}_{CE}(\hat{\boldsymbol{y}}^{\mathcal{C}_{\mathcal{M}_t}}, \boldsymbol{y}^{\mathcal{C}_{\mathcal{M}_t}})\cdot\mathbb{1}[y\notin\mathcal{C}_t]\Big], \quad (12)$$

where $\hat{\boldsymbol{y}}^{\mathcal{C}_t} = \hat{\boldsymbol{y}}_{[|\mathcal{C}_{\mathcal{M}_t}|+1:|\mathcal{C}_t'|]}$, $\hat{\boldsymbol{y}}^{\mathcal{C}_{\mathcal{M}_t}} = \hat{\boldsymbol{y}}_{[1:|\mathcal{C}_{\mathcal{M}_t}|]}$. $\boldsymbol{y}^{\mathcal{C}_t}$ and $\boldsymbol{y}^{\mathcal{C}_{\mathcal{M}_t}}$ are the one-hot vector of y in $\mathbb{R}^{|\mathcal{C}_{\mathcal{M}_t}|+1:|\mathcal{C}_t'|}$ and $\mathbb{R}^{|\mathcal{C}_{\mathcal{M}_t}|}$ respectively.

The last part of our final optimization objective is the Task-wise Knowledge Distillation (TKD) [2, 6, 50], which is used for preserving the task-wise knowledge and preventing the learned knowledge from being biased by other tasks:

$$\mathcal{L}_{TKD} = \mathbb{E}_{(\boldsymbol{x}^a, \boldsymbol{x}^v)\sim\mathcal{D}_t'}\left[\sum_{s=1}^{t} KL(\hat{\boldsymbol{y}}_t^{\mathcal{C}_s}||\hat{\boldsymbol{y}}_{t-1}^{\mathcal{C}_s})\right], \quad (13)$$

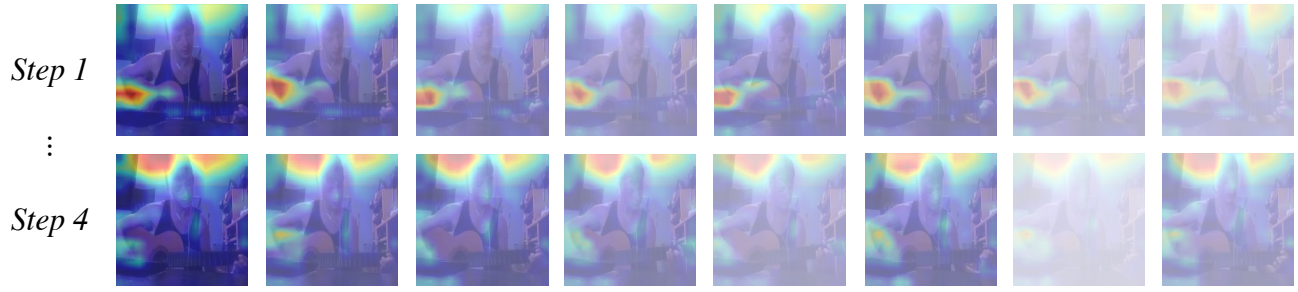


Figure 3: Visualization of the vanishing of visual attention as the incremental step increases. With the classes increasing, previously learned audio-guided visual attentive abilities, as well as the learned audio-visual correlations, are getting forgotten.

where $\hat{\boldsymbol{y}}_t^{\mathcal{C}_s}$ denotes scores of classes in $\mathcal{C}_s$ of $\hat{\boldsymbol{y}}_t$. Finally, our overall loss function is denoted as:

$$\mathcal{L}_{AV-CIL} = \mathcal{L}_{SS-CE} + \mathcal{L}_{TKD} + \mathcal{L}_{D-AVSC} + \mathcal{L}_{VAD}. \tag{14}$$

3.6. Management of Exemplar Set

In our proposed method, we hold the memory with a fixed maximum number of exemplars. In each task, we follow [2] and randomly select exemplars for new classes, and randomly reduce the number of exemplars in old classes to keep the same exemplars number in all classes.

4. Experiments

In this section, we first introduce our experimental setup, including datasets, baseline methods, and hyperparameters. Then, we show the experimental results of our proposed methods compared to the state-of-the-art methods.

Datasets. We conduct experiments with our AV-CIL compared to state-of-the-art baselines on the AVE [82], Kinetics-Sounds [3], and VGGSound [16] datasets. The AVE dataset consists of 4K 10-seconds videos from 28 audio-visual event classes. In our class-incremental settings, we randomly split the 28 audio-visual event classes into 4 incremental tasks, each of which contains 7 classes. We name it as *AVE-Class-Incremental (AVE-CI)*. Kinetics-Sounds (K-S), which contains around 24K 10-seconds videos (20K, 2K and 2K for training, validation and testing respectively) from 31 human action classes, is a subset selected from Kinetics-400 [42] dataset. In our settings, we randomly select 30 classes from the K-S and randomly divide them into 5 incremental steps, each of which contains 6 classes. We name our new constructed dataset as *Kinetics-Sounds-Class-Incremental (K-S-CI)*, which contains around 23K samples in total. The VGGSound [16] dataset contains around 200K 10-seconds videos from 309 classes. We randomly select 100 classes from the original VGGSound dataset to construct a subset named VGGSound100, which contains 60K samples in total. For each class of the VGGSound100, We randomly select 50 samples for validation and 50 samples for testing. In our class-incremental settings, we randomly divide the 100 classes into 10 incremental steps, each of which contains 10 classes. We name it as *VGGSound100-Class-Incremental (VS100-CI)*.

Baselines. We compare our proposed method with following representative and state-of-the-art methods: Fine-tuning, LwF [50], iCaRL [74], SS-IL [2], and AFC [41]. Fine-tuning is the simplest incremental training strategy that initialized the model with the parameters trained from the last task and re-train it on the current task without any constraints or other strategies to prevent the model from catastrophic forgetting. For iCaRL [74], we report the experimental results with both nearest-mean-of-exemplars (NME) [74] classification strategy and the classifier. We name them as iCaRL-NME and iCaRL-FC, respectively. Note that, iCaRL-FC is also named as iCaRL-CNN in other papers [41]. For AFC [41], we also report the experimental results with both NME classification strategy and the classifier, which are named AFC-NME and AFC-LSC respectively in our paper. We also present the experimental results of the Oracle/Upper Bound, which is defined as using all training data from seen classes to train the model. Please note that, for fair comparisons, all baselines use the same backbone as our method, including the visual encoder, audio encoder, audio-guided visual attention layer, and audio-visual fusion layer. Moreover, for exemplar-based methods iCaRL [74], SS-IL [2], and AFC [41], the exemplar selection strategies keep same as per their original papers, and all of them use the same memory size as ours.

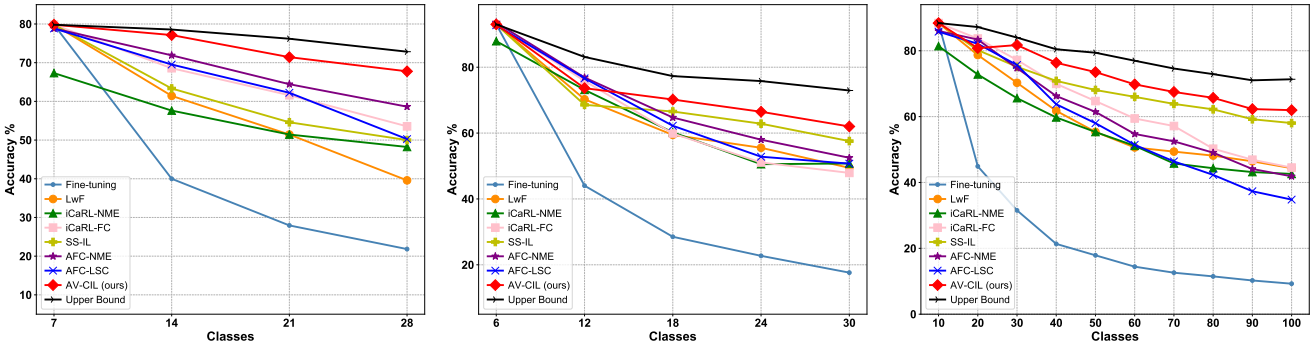
Evaluation Metric. We evaluate the performance of all the methods in our experiments with Mean Accuracy and Average Forgetting, two common evaluation metrics in class-incremental learning. Mean Accuracy is the average of the testing accuracy of each task, which can be denoted as:

$$MeanAcc. = \frac{1}{T}\sum_{t=1}^{T} a_t, \tag{15}$$

where a_t denotes testing accuracy of all seen classes after completing the training on task $\mathcal{T}_t$ (incremental step t).

Table 1: Audio-visual class-incremental results of different approaches on the AVE-CI, K-S-CI, and VS100-CI datasets. The bold part denotes the overall best results, and the underlined part denotes the best results of baselines. Our AV-CIL achieves the best performance on all three datasets.

Methods	AVE-CI		K-S-CI		VS100-CI	
	Mean Acc.	Ave. Forget.	Mean Acc.	Ave. Forget.	Mean Acc.	Ave. Forget.
Fine-tuning	42.40	70.99	41.18	89.62	26.21	89.37
LwF	58.07	26.90	65.54	16.55	59.34	23.01
iCaRL-NME	56.15	<u>11.71</u>	64.51	18.70	56.19	12.80
iCaRL-FC	65.88	26.08	65.54	40.57	64.22	29.94
SS-IL	61.94	22.49	<u>69.71</u>	<u>10.53</u>	<u>69.20</u>	<u>9.75</u>
AFC-NME	<u>68.46</u>	14.18	69.13	27.30	61.41	23.30
AFC-LSC	65.21	28.11	67.02	33.56	57.76	29.64
AV-CIL (Ours)	**74.04**	**7.63**	**73.06**	**6.48**	**72.80**	**5.49**
Oracle (Upper Bound)	76.85	–	80.43	–	78.63	–



(a) (b) (c)

Figure 4: Testing accuracy at each incremental step on (a) AVE-CI, (b) K-S-CI, and (c) VS100-CI datasets. The results show that as the incremental step increases, our AV-CIL generally outperforms other state-of-the-art incremental learning methods.

Average Forgetting is used to measure the extent of catastrophic forgetting over previously learned tasks, which can be denoted as:

$$Ave.Forget. = \frac{1}{T-1}\sum_{t=2}^{T} F_t,$$
$$s.t.\ \ F_t = \frac{1}{t-1}\sum_{i=1}^{t-1} \max_{\tau\in\{1,\dots,t-1\}} (a_{\tau,i} - a_{t,i}), \tag{16}$$

where $a_{\tau,i}$ is the testing accuracy of the i-th task after training on the τ-th task (task $\mathcal{T}_\tau$ or incremental step τ).

Implementation Details. We conduct all our experiments with PyTorch [71]. For the data pre-processing, we follow the protocol of VideoMAE [85] and AudioMAE [36]. We freeze the self-supervised pre-trained visual and audio encoder during the training of the baselines and our method, and fine-tune the rest parts of the model (audio-guided visual attention layer, audio-visual fusion layer, and the classifier). We use Adam [44] optimizer to train the model with learning rate and weight decay of 1e-3 and 1e-4, respectively. For all three datasets, we set the maximum training epochs number in each incremental step, the balance weight λ_{VAD} in $\mathcal{L}_{VAD}$, and the temperature τ in D-AVSC to 200, 0.5, and 0.05, respectively. For the AVE-CI dataset, we set the memory size to 340. For the weights of losses λ_I and λ_C, we set them to 0.5 and 1.0, respectively. For the K-S-CI dataset, we set the memory size, λ_I, and λ_C to 500, 0.1, and 1.0, respectively. For the VS100-CI dataset, the memory size, each incremental step's training epoch, λ_I, and λ_C are set to 1500, 200, 0.1, and 1.0, respectively.

4.1. Experimental Comparison

We show our main experimental results in Table 1. We can see that our proposed AV-CIL outperforms recent state-of-the-art methods significantly. Specifically, on the AVE-CI dataset, our AV-CIL outperforms the state-of-the-art Mean Accuracy and Average Forgetting results by **5.58** and **4.08**, respectively. For K-S-CI, our method outperforms the

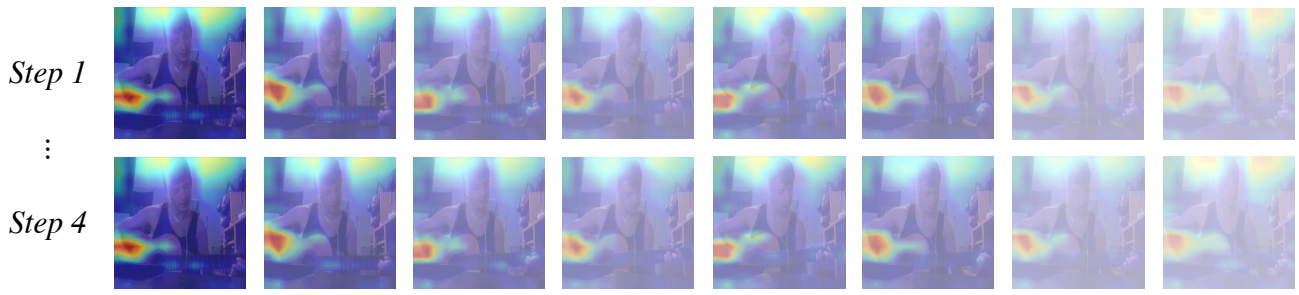


Figure 5: Visualization of the visual attention map as the incremental step increases with our proposed VAD. After the utilization of the VAD, the previously learned audio-guided visual attentive abilities, as well as the learned audio-visual correlations, are preserved by the model in new tasks.

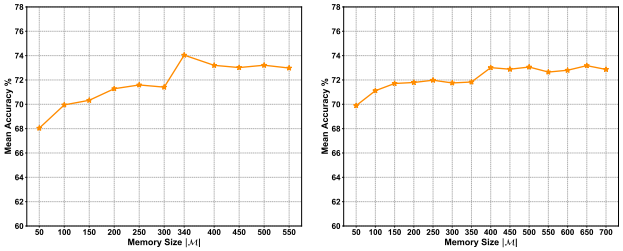
state-of-the-art method by **3.35** and **4.05** for Mean Accuracy and Average Forgetting, respectively. For the VS100-CI dataset, our method has the improvement of **3.60** and **4.26** for Mean Accuracy and Average Forgetting over state-of-the-art results. These experimental results demonstrate the effectiveness of our proposed method in audio-visual class-incremental learning. We also show the testing accuracy at each incremental step of our AV-CIL, baselines, and the upper bound on the AVE-CI, K-S-CI, and VS100-CI datasets in Figure 4a, 4b, and 4c. We can see that our method achieves the best performance at each incremental step on the AVE-CI dataset. For the K-S-CI and VS100-CI datasets, our proposed AV-CIL also has the best overall results and the best performance at each incremental step, except at step 2. In summary, our method has a significant superiority in audio-visual class-incremental learning compared to state-of-the-art class-incremental learning methods, which demonstrates the effectiveness of the combination of our proposed D-AVSC and VAD.

4.2. Ablation Study

Our proposed method AV-CIL mainly contributes two parts: Dual-Audio-Visual Similarity Constraint (D-AVSC) that consists of the Instance-aware Audio-Visual Semantic Similarity (I-AVSS) and the Class-aware Audio-Visual Semantic Similarity (C-AVSS), and Visual Attention Distillation (VAD). In this subsection, we conduct the ablation study of our method by removing single or multiple parts from D-AVSC and VAD when running experiments. The results are presented in Table 3, from which we can see that each of our proposed parts has a positive impact on the final results. This demonstrates the effectiveness of our proposed D-AVSC (including both I-AVSS and D-AVSS) and VAD.

4.3. Comparison with Single Modality Results

To evaluate the effectiveness of class-incremental learning with joint audio-visual modalities compared to using only the audio or visual modality, in this subsection, we conduct experiments of class-incremental learning with only single audio or visual modality and compare them with our experimental results with joint audio-visual modalities. We show the results in Table 2, from which we can see that, for all the baselines, compared to training with single audio or visual modality, training with audio-visual modalities achieves higher Mean Accuracy in class-incremental learning, which demonstrates that joint audio-visual modeling is better for class-incremental learning compared to only using data from single audio or visual modality for class-incremental learning. Please see Appendix for the comparison at each step.



(a) (b)

Figure 6: Experimental results of our proposed AV-CIL with different memory size ($|\mathcal{M}|$) on (a) AVE-CI, and (b) K-S-CI dataset.

4.4. Effect of Memory Size

In this section, we investigate the effect of memory size on our proposed class-incremental learning approach. We conduct experiments with different memory sizes and report the results in Figure 6a and 6b for AVE-CI and K-S-CI datasets, respectively. Our experimental results show that the performance of our model generally improves followed by stable fluctuations as the memory size increases on both datasets. Specifically, on the AVE-CI dataset, the model's performance consistently improves as the memory size grows from 50 to 340. However, a fluctuation in the

Table 2: Ablation study on input modalities. We can see that joint audio-visual modeling is better than individual modeling in class-incremental learning.

Methods	Mean Accuracy								
	AVE-CI			K-S-CI			VS100-CI		
	Audio	Visual	Aud.-Vis.	Audio	Visual	Aud.-Vis.	Audio	Visual	Aud.-Vis.
Fine-tuning	38.41	35.68	42.40	32.60	38.49	41.18	24.40	23.06	26.21
LwF [50]	51.88	50.09	58.07	47.68	61.91	65.54	49.49	49.10	59.34
iCaRL-NME [74]	44.62	52.72	56.15	39.75	61.47	64.51	41.61	46.61	56.62
iCaRL-FC [74]	56.17	55.06	65.88	43.14	58.74	65.05	54.46	48.39	64.22
SS-IL [2]	53.81	52.10	61.94	47.65	64.55	69.71	57.88	53.86	69.20
AFC-NME [41]	55.28	58.67	68.46	46.71	66.63	69.13	49.91	54.50	61.10
AFC-LSC [41]	56.36	57.92	65.21	39.89	64.88	67.02	44.82	51.69	58.15
Oracle (Upper Bound)	63.66	61.55	76.85	53.39	71.94	80.43	66.22	61.75	78.63

Table 3: Ablation study on our AV-CIL. Our full model can better handle catastrophic forgetting and achieve the best incremental learning performance on AVE-CI dataset.

	D-AVSC		VAD	Mean Acc.
	I-AVSS	C-AVSS		
AV-CIL	✗	✗	✗	62.15
	✓	✗	✗	68.38
	✗	✓	✗	63.81
	✗	✗	✓	63.49
	✓	✓	✗	69.01
	✓	✗	✓	72.35
	✗	✓	✓	63.95
	✓	✓	✓	74.04

performance is following as the memory size increases from 340 to 550. Similarly, on the K-S-CI dataset, the model's performance tends to increase as the memory size grows from 50 to 400, followed by a stable fluctuation as the memory size increases from 400 to 700. These trends suggest that the model's performance steadily improves in the early stages of memory growth, and then reaches a plateau where further increases in memory size do not lead to significant performance improvements.

4.5. Visualization of Visual Attention Map

In Figure 5, we show the visualization of the visual attention map after applying our proposed Visual Attention Distillation (VAD). The illustrated sample is the same as that in Figure 3. We can see that the attentive abilities learned in previous tasks can be preserved. This demonstrates that our proposed VAD can prevent the model from forgetting the learned audio-visual correlations in previous classes. For the full version of Figure 5, please see Appendix for details.

5. Conclusion

In this paper, we explore the effectiveness of training with joint audio-visual modalities in the alleviation of the catastrophic forgetting problem in class-incremental learning and propose the audio-visual class-incremental learning, a novel class-incremental learning problem with audio-visual data. To tackle our proposed problem, we propose a method named AV-CIL, which contains the Dual-Audio-Visual Similarity Constraint to preserve the cross-modal similarity from the perspective of both instance and class. Moreover, to preserve the learned semantic correlations between audio and visual modalities as the incremental step increases, we propose Visual Attention Distillation to distill the learned audio-guided visual attentive ability into new incremental steps. Experimental results on three constructed audio-visual class-incremental datasets AVE-CI, K-S-CI, and VS100-CI show that our proposed approach outperforms state-of-the-art methods significantly. This paper provides a new direction in class-incremental learning.

Acknowledgments. We would like to thank the anonymous reviewers for their constructive comments. This work was supported in part by gifts from Cisco Systems and Adobe. The article solely reflects the opinions and conclusions of its authors but not the funding agents.

References

[1] Hongjoon Ahn, Sungmin Cha, Donggyu Lee, and Taesup Moon. Uncertainty-based continual learning with adaptive regularization. *Advances in Neural Information Processing Systems*, 32, 2019. 2, 3

[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*, pages 824–833, 2021. 2, 3, 4, 5, 6, 9

[3] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 609–617, 2017. 2, 6

[4] Ari Benjamin, David Rolnick, and Konrad Kording. Measuring and regularizing networks in function space. In *International Conference on Learning Representations (ICLR)*, 2019. 2, 3

[5] David A Bulkin and Jennifer M Groh. Seeing sounds: visual and auditory interactions in the brain. *Current Opinion in Neurobiology*, 16(4):415–419, 2006. 1

[6] Francisco M Castro, Manuel J Marín-Jiménez, Nicolás 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. 2, 3, 5

[7] Fabio Cermelli, Dario Fontanel, Antonio Tavera, Marco Ciccone, and Barbara Caputo. Incremental learning in semantic segmentation from image labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4371–4381, 2022. 2, 3

[8] Sungmin Cha, YoungJoon Yoo, Taesup Moon, et al. Ssul: Semantic segmentation with unknown label for exemplar-based class-incremental learning. *Advances in Neural Information Processing Systems*, 34:10919–10930, 2021. 2, 3

[9] Arslan Chaudhry, Puneet K Dokania, Thalaiyasingam Ajanthan, and Philip HS Torr. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 532–547, 2018. 2, 3

[10] Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with a-GEM. In *International Conference on Learning Representations*, 2019. 2, 3

[11] 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. 2, 3

[12] Changan Chen, Ziad Al-Halah, and Kristen Grauman. Semantic audio-visual navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15516–15525, 2021. 3

[13] Changan Chen, Unnat Jain, Carl Schissler, Sebastia Vicenc Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. Soundspaces: Audio-visual navigation in 3d environments. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 17–36. Springer, 2020. 3

[14] Changan Chen, Sagnik Majumder, Ziad Al-Halah, Ruohan Gao, Santhosh Kumar Ramakrishnan, and Kristen Grauman. Learning to set waypoints for audio-visual navigation. In *International Conference on Learning Representations*, 2021. 3

[15] 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*, pages 16867–16876, 2021. 1, 2

[16] 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. 2, 6

[17] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607. PMLR, 2020. 3

[18] Ziyang Chen, David F Fouhey, and Andrew Owens. Sound localization by self-supervised time delay estimation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVI*, pages 489–508. Springer, 2022. 1, 2

[19] Prithviraj Dhar, Rajat Vikram Singh, Kuan-Chuan Peng, Ziyan Wu, and Rama Chellappa. Learning without memorizing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5138–5146, 2019. 2, 3

[20] Na Dong, Yongqiang Zhang, Mingli Ding, and Gim Hee Lee. Bridging non co-occurrence with unlabeled in-the-wild data for incremental object detection. *Advances in Neural Information Processing Systems*, 34:30492–30503, 2021. 2, 3

[21] Arthur Douillard, Yifu Chen, Arnaud Dapogny, and Matthieu Cord. Plop: Learning without forgetting for continual semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4040–4050, 2021. 2, 3

[22] Arthur Douillard, Matthieu Cord, Charles Ollion, Thomas Robert, and Eduardo Valle. Podnet: Pooled outputs distillation for small-tasks incremental learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pages 86–102. Springer, 2020. 2, 3

[23] Tao Feng, Mang Wang, and Hangjie Yuan. Overcoming catastrophic forgetting in incremental object detection via elastic response distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9427–9436, 2022. 2, 3

[24] Enrico Fini, Victor G Turrisi Da Costa, Xavier Alameda-Pineda, Elisa Ricci, Karteek Alahari, and Julien Mairal. Self-supervised models are continual learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9621–9630, 2022. 2, 3

[25] Chuang Gan, Deng Huang, Hang Zhao, Joshua B Tenenbaum, and Antonio Torralba. Music gesture for visual sound separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10478–10487, 2020. 1, 2

[26] 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

[27] Siavash Golkar, Michael Kagan, and Kyunghyun Cho. Continual learning via neural pruning. *arXiv preprint arXiv:1903.04476*, 2019. 2, 3

[28] Yuan Gong, Andrew Rouditchenko, Alexander H. Liu, David Harwath, Leonid Karlinsky, Hilde Kuehne, and James R. Glass. Contrastive audio-visual masked autoencoder. In *The Eleventh International Conference on Learning Representations*, 2023. 3

[29] Dipam Goswami, René Schuster, Joost van de Weijer, and Didier Stricker. Attribution-aware weight transfer: A warm-start initialization for class-incremental semantic segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3195–3204, 2023. 2, 3

[30] Yunhui Guo, Mingrui Liu, Tianbao Yang, and Tajana Rosing. Improved schemes for episodic memory-based lifelong learning. *Advances in Neural Information Processing Systems*, 33:1023–1035, 2020. 3

[31] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 3

[32] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 3

[33] 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*, pages 831–839, 2019. 2, 3

[34] Di Hu, Feiping Nie, and Xuelong Li. Deep multimodal clustering for unsupervised audiovisual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9248–9257, 2019. 1, 2

[35] 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*, pages 10483–10492, 2022. 1, 2

[36] Po-Yao Huang, Hu Xu, Juncheng B Li, Alexei Baevski, Michael Auli, Wojciech Galuba, Florian Metze, and Christoph Feichtenhofer. Masked autoencoders that listen. In *Advances in Neural Information Processing Systems*, 2022. 3, 4, 7

[37] Ching-Yi Hung, Cheng-Hao Tu, Cheng-En Wu, Chien-Hung Chen, Yi-Ming Chan, and Chu-Song Chen. Compacting, picking and growing for unforgetting continual learning. *Advances in Neural Information Processing Systems*, 32, 2019. 2, 3

[38] Joel Jang, Seonghyeon Ye, Sohee Yang, Joongbo Shin, Janghoon Han, Gyeonghun KIM, Stanley Jungkyu Choi, and Minjoon Seo. Towards continual knowledge learning of language models. In *International Conference on Learning Representations (ICLR)*, 2022. 2, 3

[39] KJ Joseph, Jathushan Rajasegaran, Salman Khan, Fahad Shahbaz Khan, and Vineeth N Balasubramanian. Incremental object detection via meta-learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9209–9216, 2021. 2, 3

[40] Sangwon Jung, Hongjoon Ahn, Sungmin Cha, and Taesup Moon. Continual learning with node-importance based adaptive group sparse regularization. *Advances in Neural Information Processing Systems*, 33:3647–3658, 2020. 2, 3

[41] Minsoo Kang, Jaeyoo Park, and Bohyung Han. Class-incremental learning by knowledge distillation with adaptive feature consolidation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16050–16059, 2022. 2, 3, 6, 9

[42] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 6

[43] Zixuan Ke, Yijia Shao, Haowei Lin, Tatsuya Konishi, Gyuhak Kim, and Bing Liu. Continual learning of language models. In *International Conference on Learning Representations (ICLR)*, 2023. 2, 3

[44] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *International Conference on Learning Representations (ICLR)*, 2015. 7

[45] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. 2, 3

[46] Kibok Lee, Kimin Lee, Jinwoo Shin, and Honglak Lee. Overcoming catastrophic forgetting with unlabeled data in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 312–321, 2019. 2, 3

[47] Sizhe Li, Yapeng Tian, and Chenliang Xu. Space-time memory network for sounding object localization in videos. In *32nd British Machine Vision Conference 2021, BMVC 2021, Online, November 22-25, 2021*, page 255. BMVA Press, 2021. 1, 2, 3, 4

[48] Tianjiao Li, Qiuhong Ke, Hossein Rahmani, Rui En Ho, Henghui Ding, and Jun Liu. Else-net: Elastic semantic network for continual action recognition from skeleton data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13434–13443, 2021. 2, 3

[49] Xilai Li, Yingbo Zhou, Tianfu Wu, Richard Socher, and Caiming Xiong. Learn to grow: A continual structure learning framework for overcoming catastrophic forgetting. In *International Conference on Machine Learning*, pages 3925–3934. PMLR, 2019. 2, 3

[50] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947, 2017. 1, 2, 3, 4, 5, 6, 9

[51] Yan-Bo Lin, Yu-Jhe Li, and Yu-Chiang Frank Wang. Dual-modality seq2seq network for audio-visual event localization. In *ICASSP 2019-2019 IEEE International Conference*

on Acoustics, Speech and Signal Processing (ICASSP), pages 2002–2006. IEEE, 2019. 1, 2

[52] 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. *Advances in Neural Information Processing Systems*, 34:11449–11461, 2021. 1, 2

[53] David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. *Advances in Neural Information Processing Systems*, 30, 2017. 2, 3

[54] Divyam Madaan, Jaehong Yoon, Yuanchun Li, Yunxin Liu, and Sung Ju Hwang. Representational continuity for unsupervised continual learning. In *International Conference on Learning Representations (ICLR)*, 2022. 2, 3

[55] Andrea Maracani, Umberto Michieli, Marco Toldo, and Pietro Zanuttigh. Recall: Replay-based continual learning in semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7026–7035, 2021. 2, 3

[56] Pratik Mazumder, Pravendra Singh, Kranti Kumar Parida, and Vinay P Namboodiri. Avgzslnet: Audio-visual generalized zero-shot learning by reconstructing label features from multi-modal embeddings. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3090–3099, 2021. 3

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

[58] Otniel-Bogdan Mercea, Thomas Hummel, A Sophia Koepke, and Zeynep Akata. Temporal and cross-modal attention for audio-visual zero-shot learning. In *European Conference on Computer Vision*, pages 488–505, 2022. 3

[59] Fei Mi, Liangwei Chen, Mengjie Zhao, Minlie Huang, and Boi Faltings. Continual learning for natural language generation in task-oriented dialog systems. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, volume EMNLP 2020, pages 3461–3474, 2020. 2, 3

[60] 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

[61] Shentong Mo and Pedro Morgado. Localizing visual sounds the easy way. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVII*, pages 218–234. Springer, 2022. 1, 2

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

[63] Shentong Mo, Jing Shi, and Yapeng Tian. DiffAVA: Personalized text-to-audio generation with visual alignment. *arXiv preprint arXiv:2305.12903*, 2023. 1

[64] Shentong Mo and Yapeng Tian. Multi-modal grouping network for weakly-supervised audio-visual video parsing. In *Advances in Neural Information Processing Systems*, 2022. 1, 2, 4

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

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

[67] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In *International conference on machine learning*, pages 2642–2651. PMLR, 2017. 3

[68] Youngmin Oh, Donghyeon Baek, and Bumsub Ham. ALIFE: Adaptive logit regularizer and feature replay for incremental semantic segmentation. In *Advances in Neural Information Processing Systems*, 2022. 2, 3

[69] Oleksiy Ostapenko, Mihai Puscas, Tassilo Klein, Patrick Jahnichen, and Moin Nabi. Learning to remember: A synaptic plasticity driven framework for continual learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11321–11329, 2019. 3

[70] Jaeyoo Park, Minsoo Kang, and Bohyung Han. Class-incremental learning for action recognition in videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13698–13707, 2021. 2, 3

[71] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 7

[72] Senthil Purushwalkam, Pedro Morgado, and Abhinav Gupta. The challenges of continuous self-supervised learning. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVI*, pages 702–721. Springer, 2022. 2, 3

[73] Tanzila Rahman, Bicheng Xu, and Leonid Sigal. Watch, listen and tell: Multi-modal weakly supervised dense event captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8908–8917, 2019. 1, 3

[74] 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*, pages 2001–2010, 2017. 1, 2, 3, 6, 9

[75] Thomas Scialom, Tuhin Chakrabarty, and Smaranda Muresan. Fine-tuned language models are continual learners. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6107–6122, 2022. 2, 3

[76] 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*, pages 4358–4366, 2018. 1, 2

[77] Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. In *Ad-*

vances in Neural Information Processing Systems, pages 2990–2999, 2017. 2, 3

[78] Tejas Srinivasan, Ting-Yun Chang, Leticia Leonor Pinto Alva, Georgios Chochlakis, Mohammad Rostami, and Jesse Thomason. CLiMB: A continual learning benchmark for vision-and-language tasks. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 2, 3

[79] Yapeng Tian, Chenxiao Guan, Justin Goodman, Marc Moore, and Chenliang Xu. Audio-visual interpretable and controllable video captioning. In *IEEE Conference on Computer Vision and Pattern Recognition workshops*, 2019. 1, 3

[80] 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*, pages 2745–2754, 2021. 1, 2

[81] Yapeng Tian, Dingzeyu Li, and Chenliang Xu. Unified multisensory perception: Weakly-supervised audio-visual video parsing. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 436–454. Springer, 2020. 1, 2

[82] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 247–263, 2018. 1, 2, 3, 4, 6

[83] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition workshops*, 2019. 1, 2

[84] Yapeng Tian and Chenliang Xu. Can audio-visual integration strengthen robustness under multimodal attacks? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5601–5611, 2021. 2, 3, 4

[85] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems*, 2022. 3, 4, 7

[86] Fu-Yun Wang, Da-Wei Zhou, Han-Jia Ye, and De-Chuan Zhan. FOSTER: feature boosting and compression for class-incremental learning. In *In Proceedings of the European Conference on Computer Vision (ECCV)*, pages 398–414, 2022. 2, 3

[87] Xin Wang, Yuan-Fang Wang, and William Yang Wang. Watch, listen, and describe: Globally and locally aligned cross-modal attentions for video captioning. *arXiv preprint arXiv:1804.05448*, 2018. 1, 3

[88] 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. 3

[89] Yue Wu, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, and Yun Fu. Large scale incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 374–382, 2019. 2, 3

[90] Yu Wu and Yi Yang. Exploring heterogeneous clues for weakly-supervised audio-visual video parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1326–1335, 2021. 1, 2

[91] Yu Wu, Linchao Zhu, Yan Yan, and Yi Yang. Dual attention matching for audio-visual event localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6292–6300, 2019. 1, 2, 3, 4

[92] Shipeng Yan, Lanqing Hong, Hang Xu, Jianhua Han, Tinne Tuytelaars, Zhenguo Li, and Xuming He. Generative negative text replay for continual vision-language pretraining. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVI*, pages 22–38. Springer, 2022. 2, 3

[93] Shipeng Yan, Jiangwei Xie, and Xuming He. Der: Dynamically expandable representation for class incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3014–3023, 2021. 2, 3

[94] Jaehong Yoon, Eunho Yang, Jeongtae Lee, and Sung Ju Hwang. Lifelong learning with dynamically expandable networks. In *International Conference on Learning Representations (ICLR)*, 2018. 2, 3

[95] Xiaofan Yu, Yunhui Guo, Sicun Gao, and Tajana Rosing. Scale: Online self-supervised lifelong learning without prior knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2483–2494, 2023. 3

[96] Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *International Conference on Machine Learning*, pages 3987–3995. PMLR, 2017. 2, 3

[97] Bowen Zhao, Xi Xiao, Guojun Gan, Bin Zhang, and Shu-Tao Xia. Maintaining discrimination and fairness in class incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13208–13217, 2020. 2, 3

[98] Hang Zhou, Xudong Xu, Dahua Lin, Xiaogang Wang, and Ziwei Liu. Sep-stereo: Visually guided stereophonic audio generation by associating source separation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pages 52–69. Springer, 2020. 1, 2

[99] Fei Zhu, Zhen Cheng, Xu-Yao Zhang, and Cheng-lin Liu. Class-incremental learning via dual augmentation. *Advances in Neural Information Processing Systems*, 34:14306–14318, 2021. 2, 3

[100] Fei Zhu, Xu-Yao Zhang, Chuang Wang, Fei Yin, and Cheng-Lin Liu. Prototype augmentation and self-supervision for incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5871–5880, 2021. 2, 3