Learning Continual Audio-Visual Sound Separation Models

Weiguo Pian[1] Yiyang Nan[2] Shijian Deng[1] Shentong Mo[3] Yunhui Guo[1] Yapeng Tian[1]
[1] The University of Texas at Dallas [2] Brown University [3] Carnegie Mellon University
{weiguo.pian,shijian.deng,yunhui.guo,yapeng.tian}@utdallas.edu
yiyang_nan@brown.edu shentonm@andrew.cmu.edu

1. Introduction

Humans can effortlessly separate and identify individual sound sources in daily experience. This skill plays a crucial role in our ability to understand and interact with the complex auditory environments that surround us. Inspired by the multisensory perception of humans, audio-visual sound separation aims to tackle this challenge by utilizing visual information to guide the separation of individual sound sources in an audio mixture. Recent advances in deep learning have led to significant progress in audio-visual sound separation [2, 4, 11, 12, 14, 16]. However, a limitation of recent studies in audio-visual domain is their focus on scenarios where all sound source classes are presently known, overlooking the potential inclusion of unknown sound source classes during inference. This oversight leads to the *catastrophic forgetting* problem [1, 7] *How to effectively leverage visual guidance to continuously separate sounds from new categories while preserving separation ability for old sound categories* remains an open question. To bridge this gap, we propose a novel approach named ContAV-Sep (**Cont**inual **A**udio-**V**isual Sound **Sep**aration) by integrating audio-visual sound separation with continual learning principles. In our ContAV-Sep framework, we introduce a novel *Cross-modal Similarity Distillation Constraint (CrossSDC)* to not only maintain the cross-modal semantic similarity through incremental tasks but also preserve previously learned knowledge of semantic similarity in old models to counter catastrophic forgetting.

2. Method

2.1. Problem Formulation

Audio-Visual Sound Separation. Audio-visual sound separation aims to separate distinctive sound signals according to the given associated visual guidance. Following previous works [2, 4, 12], we adopt the common "mix-and-separation" training strategy to train the model. Given two videos $\boldsymbol{V}_1(\boldsymbol{s}_1, \boldsymbol{v}_1)$ and $\boldsymbol{V}_2(\boldsymbol{s}_2, \boldsymbol{v}_2)$, we can obtain the input mixed sound signal $\boldsymbol{S}$ by mixing two video sound signals $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$, and then we can have the ratio masks $\boldsymbol{mask}^1 = \boldsymbol{s}_1/\boldsymbol{S}$ and $\boldsymbol{mask}^2 = \boldsymbol{s}_2/\boldsymbol{S}$ The goal of the task is to utilize the corresponding visual guidance $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ to predict the ratio masks for reconstructing the two individual audio signals. And then, the original sound signals $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$ are used to calculate the loss function for optimizing the model:

$$\boldsymbol{\Theta}^* = \underset{\boldsymbol{\Theta}}{\operatorname{argmin}}\, \mathbb{E}_{(\boldsymbol{V}_1,\boldsymbol{V}_2)\sim\mathcal{D}}\Big[\mathcal{L}(\hat{\boldsymbol{mask}}^1, \boldsymbol{mask}^1) + \mathcal{L}(\hat{\boldsymbol{mask}}^2, \boldsymbol{mask}^2)\Big], \tag{1}$$

where $\mathcal{D}$ denotes the training set, and $\mathcal{L}$ is the loss function between the prediction and ground-truth.

Continual Audio-Visual Sound Separation. Our proposed continual audio-visual sound separation aims to train a model $\mathcal{F}_{\boldsymbol{\Theta}}$ continually on a sequence of T separation tasks $\{\mathcal{T}_1, \mathcal{T}_2, ..., \mathcal{T}_T\}$. For the t-th task $\mathcal{T}_t$ (incremental step t), we have a training set $\mathcal{D}_t = \{\boldsymbol{V}^i(\boldsymbol{s}^i, \boldsymbol{v}^i), y_t^i\}_{i=1}^{n_t}$, where i and n_t denote the i-th video sample and the total number of samples in $\mathcal{D}_t$ respectively, and $y_t^i \in \mathcal{C}_t$ is the corresponding sound source class of video $\boldsymbol{V}^i$, where $\mathcal{C}_t$ is the training sound class label space of task $\mathcal{T}_t$. For any two tasks $\mathcal{T}_{t_1}$ and $\mathcal{T}_{t_2}$ and their corresponding training sound class label space $\mathcal{C}_{t_1}$ and $\mathcal{C}_{t_2}$, we have $\mathcal{C}_{t_1} \cap \mathcal{C}_{t_2} = \emptyset$. For a task $\mathcal{T}_t$, where $t > 1$, holding a small size of memory/exemplar set $\mathcal{M}_t$ to store some data from old tasks is permitted in our setting. Therefore, with the memory/exemplar set, all available data that can be used for training in task $\mathcal{T}_t$ ($t > 1$) can be denoted as $\mathcal{D}'_t = \mathcal{D}_t \cup \mathcal{M}_t$. Finally, the training process of Eq. 1 in our continual audio-visual sound separation setting can be denoted as:

$$\boldsymbol{\Theta}_t = \underset{\boldsymbol{\Theta}_{t-1}}{\operatorname{argmin}}\, \mathbb{E}_{(\boldsymbol{V}_1,\boldsymbol{V}_2)\sim\mathcal{D}'_t}\Big[\mathcal{L}(\hat{\boldsymbol{mask}}^1, \boldsymbol{mask}^1) + \mathcal{L}(\hat{\boldsymbol{mask}}^2, \boldsymbol{mask}^2)\Big],$$
$$s.t.\ \ \hat{\boldsymbol{mask}}^1 = \mathcal{F}_{\boldsymbol{\Theta}_{t-1}}(\boldsymbol{S}, \boldsymbol{v}_1), \hat{\boldsymbol{mask}}^2 = \mathcal{F}_{\boldsymbol{\Theta}_{t-1}}(\boldsymbol{S}, \boldsymbol{v}_2), \tag{2}$$

which means that, the new model $\boldsymbol{\Theta}_t$ is obtained by updating the old model $\boldsymbol{\Theta}_{t-1}$ which was trained on the previous task, using current task's available data $\mathcal{D}'_t$.

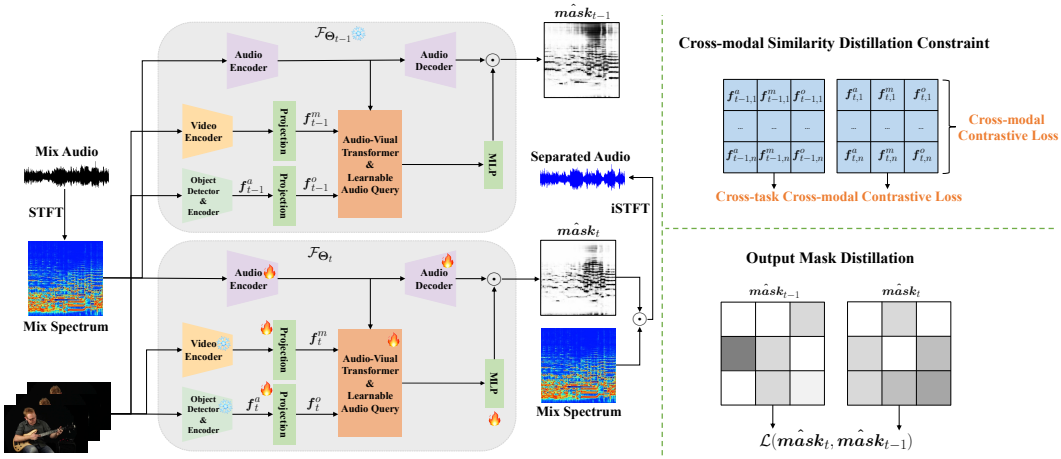

Figure 1. Overview of our proposed ContAV-Sep

2.2. Overview

An overview of our proposed ContAV-Sep is illustrated in Fig. 1. The ContAV-Sep consists of a separation base model, an output mask distillation module, and our proposed *Cross-modal Similarity Distillation Constraint (CrossSDC)*. We use the state-of-the-art audio-visual separator: iQuery [2] as the base model of our approach, which contains a video encoder to extract the global motion feature, an object detector and image encoder to obtain the object feature, a U-Net [10] for mixture sound encoding and separated sound decoding, and an audio-visual Transformer to get the separated sound feature through multi-modal cross-attention mechanism and class-aware audio queries.

2.3. Cross-modal Similarity Distillation Constraint

According to [8, 9], cross-modal semantic correlation in audio-visual modeling tends to diminish during subsequent incremental phases, which leads to catastrophic forgetting in our continual audio-visual sound separation task. We propose a novel Cross-modal Similarity Distillation Constraint (CrossSDC) that serves two crucial purposes (1) maintaining cross-modal semantic similarity through incremental tasks, and (2) preserve previous learned semantic similarity knowledge from old tasks.

CrossSDC incorporates the cross-modal similarity knowledge acquired from previous tasks into the contrastive loss. This integration not only facilitates the learning of cross-modal semantic similarities in new tasks but also ensures the preservation of previously acquired knowledge. In the incremental step t ($t > 1$), the instance-aware part of our CrossSDC can be formulated as:

$$\mathcal{L}_{inst.} = -\mathbb{E}_{\boldsymbol{V}^i \sim \mathcal{D}'_t} \left[\frac{1}{\sum_j \mathbb{1}[i=j]} \sum_j \mathbb{1}[i=j] \log \frac{\exp(\text{sim}(\boldsymbol{f}^{mod_1}_{\tau,i}, \boldsymbol{f}^{mod_2}_{\tau,j}))}{\sum_j \exp(\text{sim}(\boldsymbol{f}^{mod_1}_{\tau,i}, \boldsymbol{f}^{mod_2}_{\tau,j}))} \right], \tag{3}$$

where $\mathbb{1}[i=j]$ is an indicator that equals 1 when $i=j$, denoting that video samples $\boldsymbol{V}^i$ and $\boldsymbol{V}^j$ are the same video; The sim function represents the cosine similarity function with temperature scaling;

To preserve the semantic similarity within each class across incremental tasks, we also incorporate a class-aware component specifically designed for inter-class cross-modal semantic similarity, which can be formulated as:

$$\mathcal{L}_{cls.} = -\mathbb{E}_{(\boldsymbol{V}^i, y^i) \sim \mathcal{D}'_t} \left[\frac{1}{\sum_j \mathbb{1}[y^i=y^j]} \sum_j \mathbb{1}[y^i=y^j] \log \frac{\exp(\text{sim}(\boldsymbol{f}^{mod_1}_{\tau,i}, \boldsymbol{f}^{mod_2}_{\tau,j}))}{\sum_j \exp(\text{sim}(\boldsymbol{f}^{mod_1}_{\tau,i}, \boldsymbol{f}^{mod_2}_{\tau,j}))} \right]. \tag{4}$$

So, visual and audio features from two videos are encouraged to be close when they belong to the same class. The overall formulation of our CrossSDC is as follows:

$$\mathcal{L}_{CrossSDC} = \lambda_{ins}\mathcal{L}_{ins} + \lambda_{cls}\mathcal{L}_{cls}, \tag{5}$$

where λ_{ins} and λ_{cls} are two scalars that balance the contributions of the two loss terms.

2.4. Overall Loss Function

To effectively combine CrossSDC with the overall objective, we incorporate it alongside output distillation and the main separation loss function.

In our approach, we distill knowledge for data from the memory set and utilize the output of the old model as the distillation target to preserve this knowledge.

$$\mathcal{L}_{dist.} = \mathbb{E}_{(\boldsymbol{V}^i_1, \boldsymbol{V}^i_2) \sim \mathcal{M}_t} \Big[||\hat{\boldsymbol{mask}}^1_t - \hat{\boldsymbol{mask}}^1_{t-1}||_1 + ||\hat{\boldsymbol{mask}}^2_t - \hat{\boldsymbol{mask}}^2_{t-1}||_1 \Big], \tag{6}$$

where $\hat{\boldsymbol{mask}}^1_{t-1}$ and $\hat{\boldsymbol{mask}}^2_{t-1}$ are predicted masks generated by the old model that is trained at incremental step $t-1$. For the loss function here, we follow [2, 15] and

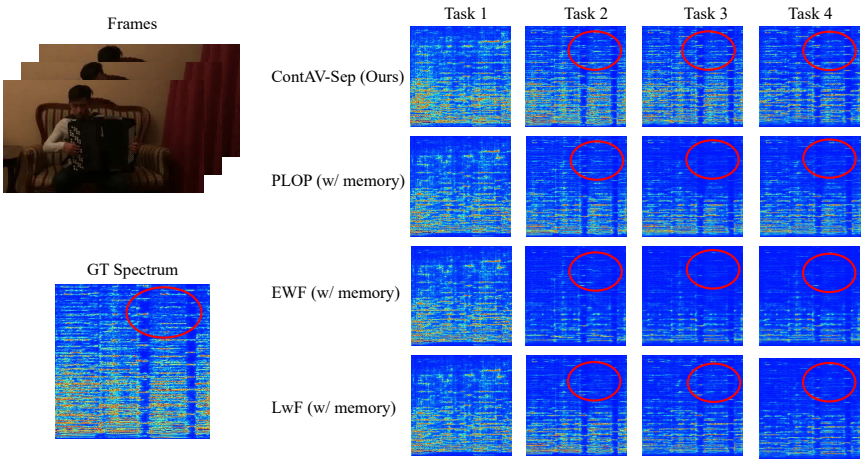

Figure 2. Left: A randomly selected sample with its frame and ground-truth spectrum. Right: Visualization of the separated sound by our ContAV-Sep and baselines at each incremental step.

adopt the per-pixel L_1 loss [15]. For the main separation loss function, we also apply the per-pixel L_1 loss:

$$\mathcal{L}_{main} = \mathbb{E}_{(\boldsymbol{V}_1^i, \boldsymbol{V}_2^i)\sim\mathcal{M}_t}\Big[||\hat{\boldsymbol{mask}}_t^1 - \boldsymbol{mask}^1||_1 + ||\hat{\boldsymbol{mask}}_t^2 - \boldsymbol{mask}^2||_1\Big], \quad (7)$$

Finally, our overall loss function is denoted as:

$$\mathcal{L}_{ContAV-Sep} = \mathcal{L}_{main} + \lambda_{dist.}\mathcal{L}_{dist.} + \mathcal{L}_{CrossSDC}, \quad (8)$$

2.5. Management of Memory Set

Our proposed framework maintains a compact memory set throughout incremental updates. We randomly select exemplars for each current class and combining these new exemplars with the existing memory set.

3. Experiments

3.1. Experimental Setup

Dataset. Follow common practice [2, 17], we conducted experiments on *MUSIC-21* [16], which contains solo videos of 21 instruments categories.

Baselines. We compare our proposed approach with vanilla Fine-tuning strategy, and continual learning methods EWC [5] and LwF [7]. We also select two state-of-the-art continual semantic segmentation methods PLOP [3] and EWF [13] as our baselines.

Evaluation. We follow previous works [2, 12] in sound separation, and evaluate the performance of all the methods using: Signal to Distortion Ratio (SDR), Signal to Interference Ratio (SIR), and Signal to Artifact Ratio (SAR). For all these three metrics, higher values denote better results.

3.2. Experimental Comparison

The main experimental comparisons are shown in Tab. 1. Our proposed method, ContAV-Sep, outperforms state-of-the-art baselines by a substantial margin. Notably, ContAV-Sep achieves a 0.3 improvement in SDR over the best compared method. Additionally, our method surpasses the top baseline by 0.25 in SIR and 0.41 in SAR.

Table 1. Main results of different methods on MUSIC-21 dataset

Method	SDR↑	SIR↑	SAR↑
w/o memory			
Fine-tuning	3.46	9.30	10.57
LwF [7]	3.45	8.78	10.66
EWC [6]	3.67	9.58	10.30
PLOP [3]	3.82	10.06	10.22
EWF [13]	3.98	9.68	11.52
w/ memory			
LwF [7]	6.76	12.77	12.60
EWC [6]	6.65	13.01	11.73
PLOP [3]	7.03	13.30	11.90
EWF [13]	5.35	11.35	11.81
ContAV-Sep (Ours)	**7.33**	**13.55**	**13.01**
Upper Bound (Oracle)	10.36	16.64	14.68

Our observations further demonstrate that retaining a small memory set significantly enhances the performance of each baseline method. For instance, equipping LwF [7] with a small memory set results in improvements of 3.31, 3.99, and 1.94 on SDR, SIR, and SAR, respectively. Our method is consistently observed to outperform others in terms of SDR at all incremental steps.

3.3. Ablation Study on CrossSDC

In this subsection, we conduct an ablation study to investigate the effectiveness of our proposed CrossSDC. By removing single or multiple components of the CrossSDC, we evaluate the impact of each on the final results. The results of the ablation study are presented in Tab. 2. From the

Table 2. Ablation study on our proposed ContAV-Sep. Our full approach achieves best results compared to the variants.

	$\mathcal{L}_{dist.}$	$\mathcal{L}_{inst.}$	$\mathcal{L}_{cls.}$	SDR↑	SIR↑	SAR↑
ContAV-Sep	✔	✗	✗	6.32	12.99	11.82
	✔	✔	✗	6.01	11.92	11.74
	✔	✗	✔	6.86	13.12	12.25
	✔	✔	✔	**7.33**	**13.55**	**13.01**

Table 3. Results of ContAV-Sep with different memory size

	# of sample per class	SDR↑	SIR↑	SAR↑
ContAV-Sep	1	7.33	13.55	13.01
	2	7.26	13.10	12.65
	3	7.88	13.66	**13.43**
	4	**8.16**	**14.16**	13.21

table, we can see that our full model achieves the best performance compared to the variants, which further demonstrates the effectiveness of our proposed CrossSDC.

3.4. Effect on Memory Size

The default setting of the memory size is 1 sample per old classes.We conduct experiments by increasing the memory size from 1 sample per old classes to 4 samples per old classes. The results are shown in Tab. 3.

3.5. Visualization of Separated Sounds

Figure 2 presents a visualization of the separated results across successive incremental steps. We highlight the area at the top right part of the spectrum.

4. Conclusion

In this paper, we explore training audio-visual sound separation models under a more practical scenario of continual learning, and propose the continual audio-visual sound separation task. To tackle our proposed new problem, we propose ContAV-Sep, which involves a Cross-modal Similarity Distillation Constraint (CrossSDC) to maintain cross-modal semantic similarity through incremental tasks, as well as preserving previous learned semantic similarity knowledge from old tasks. Experimental results on the MUSIC-21 dataset demonstrate the superiority of our method in our proposed continual audio-visual sound separation task.

Acknowledgments. This work was supported in part by a Cisco Faculty Research Award, an Amazon Research Award, and a research gift from Adobe. The article solely reflects the opinions and conclusions of its authors but not the funding agents.

References

[1] Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *ECCV*, 2018. 1

[2] Jiaben Chen, Renrui Zhang, Dongze Lian, Jiaqi Yang, Ziyao Zeng, and Jianbo Shi. iquery: Instruments as queries for audio-visual sound separation. In *CVPR*, 2023. 1, 2, 3

[3] Arthur Douillard, Yifu Chen, Arnaud Dapogny, and Matthieu Cord. Plop: Learning without forgetting for continual semantic segmentation. In *CVPR*, 2021. 3

[4] Ruohan Gao and Kristen Grauman. Co-separating sounds of visual objects. In *ICCV*, 2019. 1

[5] 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. *PNAS*, 2017. 3

[6] Kibok Lee, Kimin Lee, Jinwoo Shin, and Honglak Lee. Overcoming catastrophic forgetting with unlabeled data in the wild. In *ICCV*, 2019. 3

[7] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE TPAMI*, 2017. 1, 3

[8] Shentong Mo, Weiguo Pian, and Yapeng Tian. Class-incremental grouping network for continual audio-visual learning. In *ICCV*, 2023. 2

[9] Weiguo Pian, Shentong Mo, Yunhui Guo, and Yapeng Tian. Audio-visual class-incremental learning. In *ICCV*, 2023. 2

[10] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 2

[11] Yiyang Su, Ali Vosoughi, Shijian Deng, Yapeng Tian, and Chenliang Xu. Separating invisible sounds toward universal audiovisual scene-aware sound separation. *arXiv preprint arXiv:2310.11713*, 2023. 1

[12] Yapeng Tian, Di Hu, and Chenliang Xu. Cyclic co-learning of sounding object visual grounding and sound separation. In *CVPR*, 2021. 1, 3

[13] Jia-Wen Xiao, Chang-Bin Zhang, Jiekang Feng, Xialei Liu, Joost van de Weijer, and Ming-Ming Cheng. Endpoints weight fusion for class incremental semantic segmentation. In *CVPR*, 2023. 3

[14] Yuxin Ye, Wenming Yang, and Yapeng Tian. Lavss: Location-guided audio-visual spatial audio separation. In *WACV*, 2024. 1

[15] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *ECCV*, 2018. 2, 3

[16] Hang Zhao, Chuang Gan, Wei-Chiu Ma, and Antonio Torralba. The sound of motions. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1735–1744, 2019. 1, 3

[17] Lingyu Zhu and Esa Rahtu. Visually guided sound source separation and localization using self-supervised motion representations. In *WACV*, 2022. 3