Continual Audio-Visual Sound Separation Supplementary Material

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

A Appendix

In this appendix, we present a more detailed model architecture in section A.1. Following that, we show the experimental comparison to the uni-modal baseline to prove the effectiveness of cross-modal similarity modeling and preservation in section A.2. After that, in section A.3, we conduct experiments on the AVE [9] and the VGGSound [1] datasets, which includes a broader range of audio-visual data beyond the music domain. Furthermore, in section A.4, we present the performance on old classes of our proposed method and baselines, to further prove that our method can better mitigate the catastrophic forgetting problem compared to baselines. Finally, we offer additional visualization results in section A.5, to further demonstrate that our method can better handle the catastrophic forgetting problem in the proposed Continual Audio-Visual Sound Separation task compared to other continual learning baseline methods.

A.1 Model Architecture

Audio Network. Following [2], we use the U-Net [8] framework as the architecture of our audio network. In our experiments, the audio network has 7 down-convolutions and 7 up-convolutions. It takes a 2D Time-Frequency spectrum with size of $1 \times 256 \times 256$ as an input to generate the latent representation with a size of $256 \times 32 \times 32$ through the encoder part. Finally, it outputs the audio embedding with a size of $32 \times 256 \times 256$ through the decoder.

Audio-Visual Transformer. We follow the implementation in [2], and use the Transformer decoder architecture as our audio-visual Transformer. The audio-visual Transformer consists of 4 decoder layers, with the first layer being the motion cross-attention layer and the following 3 layers performing audio cross-attention and self-attention operation. The audio-visual Transformer generates the separated audio feature with a dimension of 256, followed by a two-layers MLP to obtain the mask embedding with a dimension of 32. Then, the channel-wise multiplication is applied between the generated mask embedding and the audio embedding to obtain the predicted mask $\hat{\boldsymbol{M}}$.

Video Encoder & **Object Detector** & **Object Encoder.** For the pre-trained video encoder, object detector, and object encoder, we use the VideoMAE [10], Detic [12], and CLIP [7], respectively. These models are frozen during the training process and the input size, out size, and internal feature dimensions remain the same as in their original implementations.

A.2 Compared to Uni-modal Baseline

To evaluate the superiority of cross-modal similarity modeling and preservation in continual audio-visual sound separation, in this subsection, we constructed a variant of our ContAV-Sep, in which we modify our proposed CrossSDC to the intra-modal similarity distillation version. We name it as ContAV-Sep-intra. The experimental results are shown in Tab. 1. We can see that, our ContAV-Sep

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

outperforms the variant significantly, further validating the effectiveness of modeling and preserving cross-modal similarity.

Table 1: Comparison to the uni-modal variant on MUSIC-21 dataset.

Methods	SDR↑	SIR↑	SAR↑
ContAV-Sep-intra	6.86	13.13	12.31
ContAV-Sep	**7.33**	**13.55**	**13.01**

A.3 Experiments on the AVE and the VGGSound datasets

To further evaluate the efficacy of our proposed method across a broader sound domain, we conduct experiments using the AVE [9] and the VGGSound [1] datasets. In the experiments on the AVE dataset, we randomly split the 28 classes in the AVE dataset into 4 tasks, each of which contains 7 classes. The results are presented in Tab. 2, in which our ContAV-Sep outperforms the baseline models in terms of the SDR and SIR metrics, further demonstrating the robustness of our method beyond the domain of musical sounds. However, it was noted that both our method and the upper bound exhibit relatively low SAR scores when compared to the baselines. Gao et al. [4] provide an interpretation for this phenomenon, explaining that the SAR primarily captures the absence of artifacts. Therefore, it can remain high even when the separation quality is suboptimal. In contrast, the SDR and SIR metrics are used to evaluate the accuracy of the separation. For the experiments on the VGGSound [1] dataset, we follow [6] and randomly selecting 100 classes for continual learning. These classes are divided into 4 tasks, each containing 25 classes. Given the significantly larger number of samples per class in the VGGSound dataset, we set the memory size to 20 samples per class for methods that utilize memory. The experimental results, shown in Tab. 3, demonstrate that our proposed ContAV-Sep consistently outperforms the baseline methods on the VGGSound dataset in the context of continual audio-visual sound separation.

Table 2: Continual audio-visual separation results on the AVE [9] dataset.

Methods	SDR↑	SIR↑	SAR↑
iQuery [2] + Fine-tuning	2.07	5.64	**12.83**
iQuery [2] + LwF (w/ memory) [5]	2.19	6.43	10.67
iQuery [2] + PLOP (w/ memory) [3]	2.45	6.11	11.57
iQuery [2] + AV-CIL (w/ memory) [6]	2.53	6.64	11.26
ContAV-Sep (Ours)	**2.72**	**7.32**	9.86
Upper Bound (with iQuery)	3.55	8.53	9.63

Table 3: Continual audio-visual separation results on the VGGSound [1] dataset.

Methods	SDR↑	SIR↑	SAR↑
iQuery [2] + Fine-tuning	3.69	7.23	**12.84**
iQuery [2] + LwF (w/ memory) [5]	4.71	8.89	11.70
iQuery [2] + PLOP (w/ memory) [3]	4.56	8.32	12.34
iQuery [2] + AV-CIL (w/ memory) [6]	4.69	8.61	11.60
ContAV-Sep (Ours)	**4.90**	**9.25**	11.73
Upper Bound (with iQuery)	5.57	9.19	13.24

A.4 Results on Old Classes

To further evaluate our proposed method's ability to handle the catastrophic forgetting problem, we use SDR_t^{old}, SIR_t^{old}, and SAR_t^{old} to denote the separation performance on old classes after training at incremental step t. And then, we average each of these three metrics over all incremental steps,

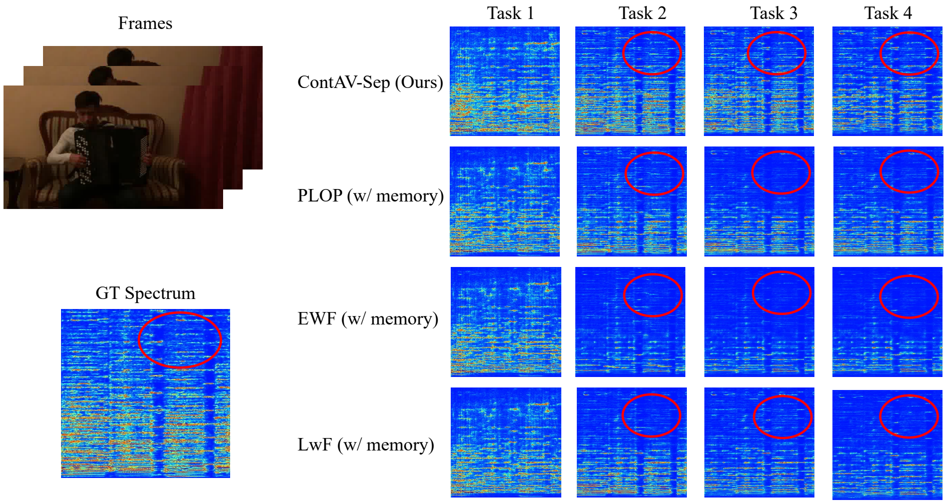

Figure 1: Left: a randomly selected sample with its frame and ground-truth spectrogram. Right: separated sounds by our ContAV-Sep and baselines at each incremental step.

yielding SDR_{mean}^{old}, SIR_{mean}^{old}, and SAR_{mean}^{old}:

$$\begin{aligned} SDR_{mean}^{old} &= \frac{1}{t-1}\sum_{i=2}^{t} SDR_{t}^{old}, \\ SIR_{mean}^{old} &= \frac{1}{t-1}\sum_{i=2}^{t} SIR_{t}^{old}, \\ SAR_{mean}^{old} &= \frac{1}{t-1}\sum_{i=2}^{t} SAR_{t}^{old}. \end{aligned} \tag{1}$$

The results are presented in Tab. 4, where it is evident that our method outperforms the baseline models, indicating a superior capability in addressing catastrophic forgetting within the context of continual audio-visual sound separation.

Table 4: Experimental results on old classes on MUSIC-21 dataset.

Methods	$SDR_{mean}^{old} \uparrow$	$SIR_{mean}^{old} \uparrow$	$SAR_{mean}^{old} \uparrow$
iQuery [2] + LwF [5] (w/ memory)	7.61	13.76	12.90
iQuery [2] + PLOP [3] (w/ memory)	7.90	14.05	12.54
iQuery [2] + EWF [11] (w/ memory)	6.96	13.20	12.86
ContAV-Sep (Ours)	**8.11**	**14.30**	**13.26**

A.5 Visualization Results

We present visualization results in Fig. 1, 2, 3, 4, 5, and 6. In each figure, we show the results of our proposed ContAV-Sep, along with four baseline methods PLOP [3], EWF [11], and Lwf [5] at each incremental step. We highlight specific areas after the final step, where it is evident that our method yields a more accurate predicted spectrum and preserves more details after training on new tasks, which demonstrates that our method can better handle the catastrophic forgetting problem compared to baseline continual learning methods.

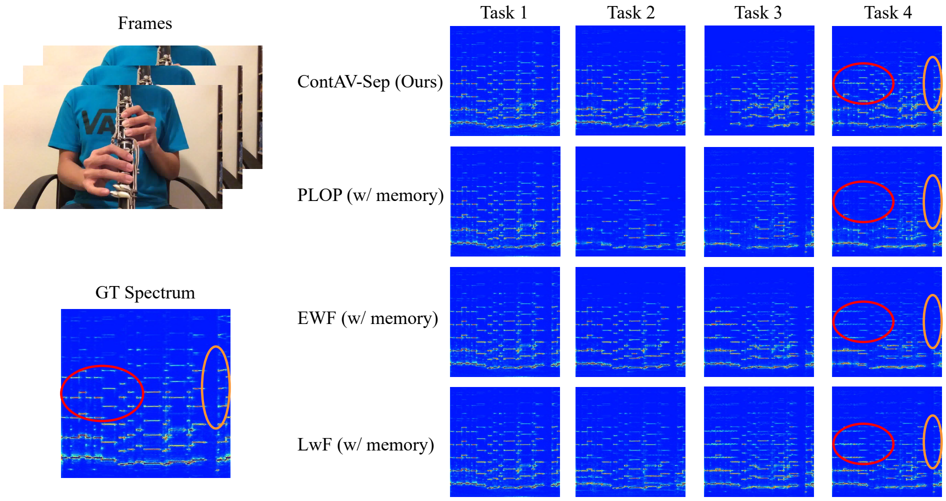

Figure 2: Left: A randomly selected sample with its frame and ground-truth spectrum. Right: Visualization of the separated sound by our ContAV-Sep and baselines at each incremental step.

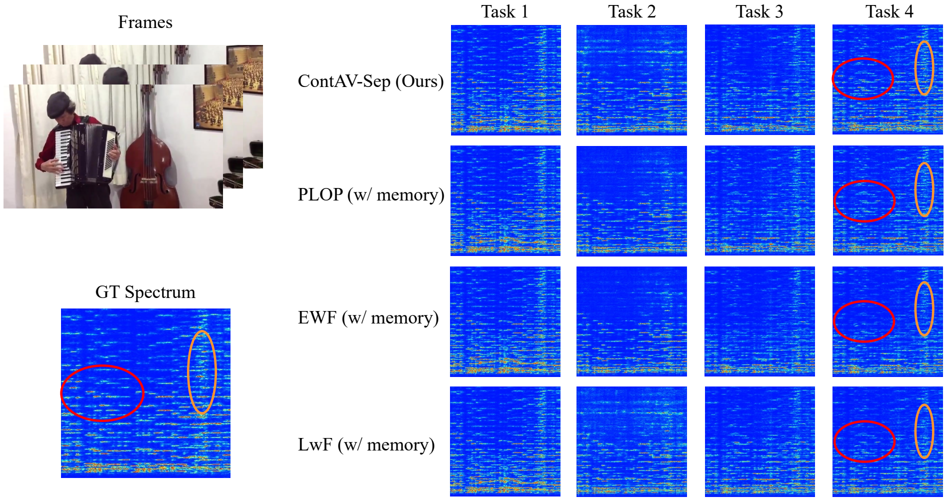

Figure 3: Left: A randomly selected sample with its frame and ground-truth spectrum. Right: Visualization of the separated sound by our ContAV-Sep and baselines at each incremental step.

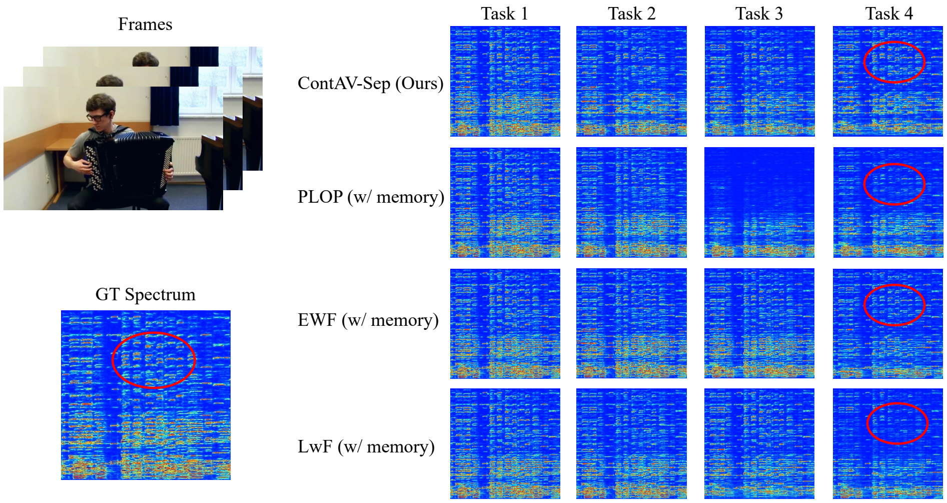

Figure 4: Left: A randomly selected sample with its frame and ground-truth spectrum. Right: Visualization of the separated sound by our ContAV-Sep and baselines at each incremental step.

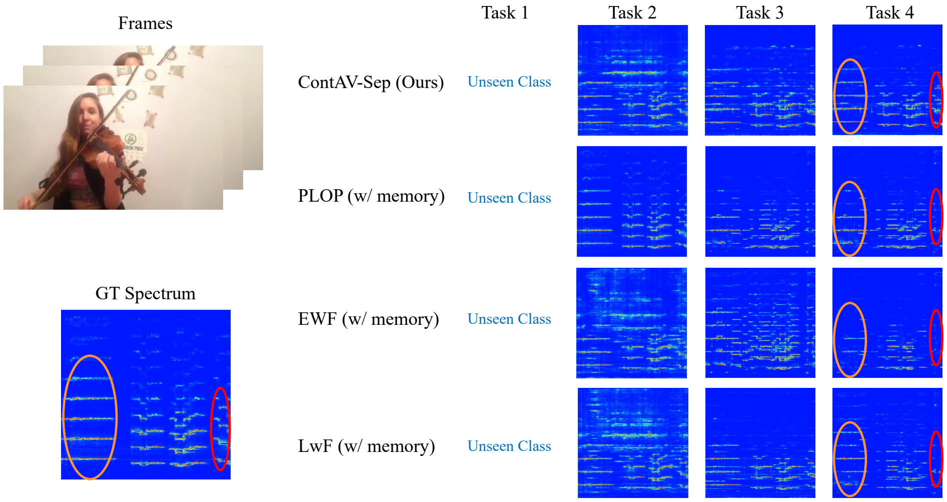

Figure 5: Left: A randomly selected sample with its frame and ground-truth spectrum. Right: Visualization of the separated sound by our ContAV-Sep and baselines at each incremental step.

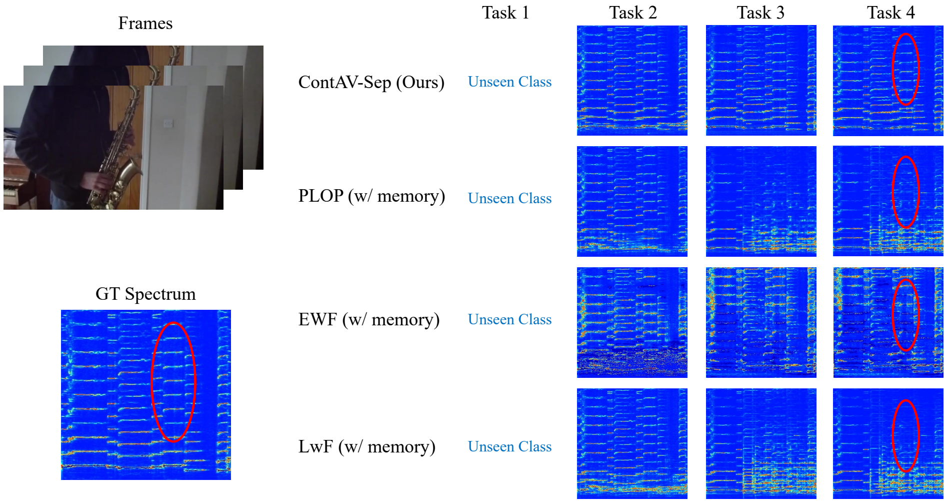

Figure 6: Left: A randomly selected sample with its frame and ground-truth spectrum. Right: Visualization of the separated sound by our ContAV-Sep and baselines at each incremental step.

References

[1] 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] Jiaben Chen, Renrui Zhang, Dongze Lian, Jiaqi Yang, Ziyao Zeng, and Jianbo Shi. iquery: Instruments as queries for audio-visual sound separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14675–14686, 2023.

[3] 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.

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

[5] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017.

[6] Weiguo Pian, Shentong Mo, Yunhui Guo, and Yapeng Tian. Audio-visual class-incremental learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7799–7811, 2023.

[7] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[8] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[9] 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.

[10] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022.

[11] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7204–7213, 2023.

[12] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *European Conference on Computer Vision*, pages 350–368. Springer, 2022.