Unveiling the Power of Audio-Visual Early Fusion Transformers with Dense Interactions through Masked Modeling
(*Supplementary Material*)

Shentong Mo
Carnegie Mellon University
shentong.mo@sv.cmu.edu

Pedro Morgado
University of Wisconsin-Madison
pmorgado@wisc.edu

https://github.com/stoneMo/DeepAVFusion

In this supplementary material, we first provide detailed implementation details on each audio-visual main downstream task and additional analysis of DeepAVFusion. We also provide a few demonstrations of sound source localization, segmentation, and separation.

A. Implementation Details

A.1. Audio-visual Classification

We conduct audio-visual classification on VGGSounds using two different downstream evaluation protocols: linear probing and fine-tuning. In both cases, we attach a linear layer to the pre-trained encoder as a classification head. Specifically, we (average) pool audio representations from all time-frequency patches into a single global audio representation, visual representations from all spatial patches into a global image representation, and the representations obtained for the fusion tokens into a global audio-visual representation. The input to the linear classifier is a subset of these three global features. In the main paper, we used only audio and visual representations. Below, we provide an analysis using different subsets that include fusion tokens. Linear probe evaluations only train the linear head to evaluate the quality of pre-trained features, while fine-tuning evaluations finetune the full model to assess DeepAVFusion's ability to provide strong initializations. In both cases, the models are trained for 50 epochs using the Adam optimizer [2] with a learning rate of $1e-4$ and batch size of 128.

A.2. Visually-Guided Sound Source Localization and Segmentation

We further evaluate our models on sound source localization and segmentation, using a supervised transfer learning protocol. We assess source localization on Flickr-SoundNet, following the work of [5]. For DeepAVFusion, the localization prediction maps $\hat{y}$ are defined as the cosine similarity between the global audio representation and the local visual representations. The full model is then trained to minimize the average per-pixel binary cross entropy to ground truth maps y using the Adam optimizer [2] for 30 epochs, with a learning rate of $1e-4$ and a batch size of 128. To compare with prior work, we use the same downstream evaluation protocol (both the prediction head and training procedure), while using the different objectives proposed in the original papers to pre-train the same ViT-Base backbones (as used by DeepAVFusion). Source segmentation is evaluated on AVSBench using the training protocol of prior work [8]. Specifically, to compute the segmentation maps, we fuse the global audio representation with the local visual representations, and upsample these localized audio-visual representations using an upsampling decoder with three sequential up-convolution blocks. The full model is then trained to minimize the binary cross entropy (BCE) between the segmentation maps predicted from the upsampling decoder and ground-truth masks. The model is trained for 20 epochs using the Adam optimizer [2] with a learning rate of $1e-4$ and a batch size of 128.

A.3. Sound Source Separation

For sound source separation, we follow the training protocol of [4, 7]. We attach an audio U-Net decoder to the pre-trained audio-visual encoders. The decoder receives the representations from an audio mixture and from the target image (containing the object whose sound should be separated). Then, through a series of transposed convolutions and an output head, the decoder predicts a time-frequency separation mask, which is used to modulate the STFT of the input mixture in order to predict the separated sound source. Similarly to [4, 7], the model is trained to minimize the binary cross-entropy between the predicted separation masks

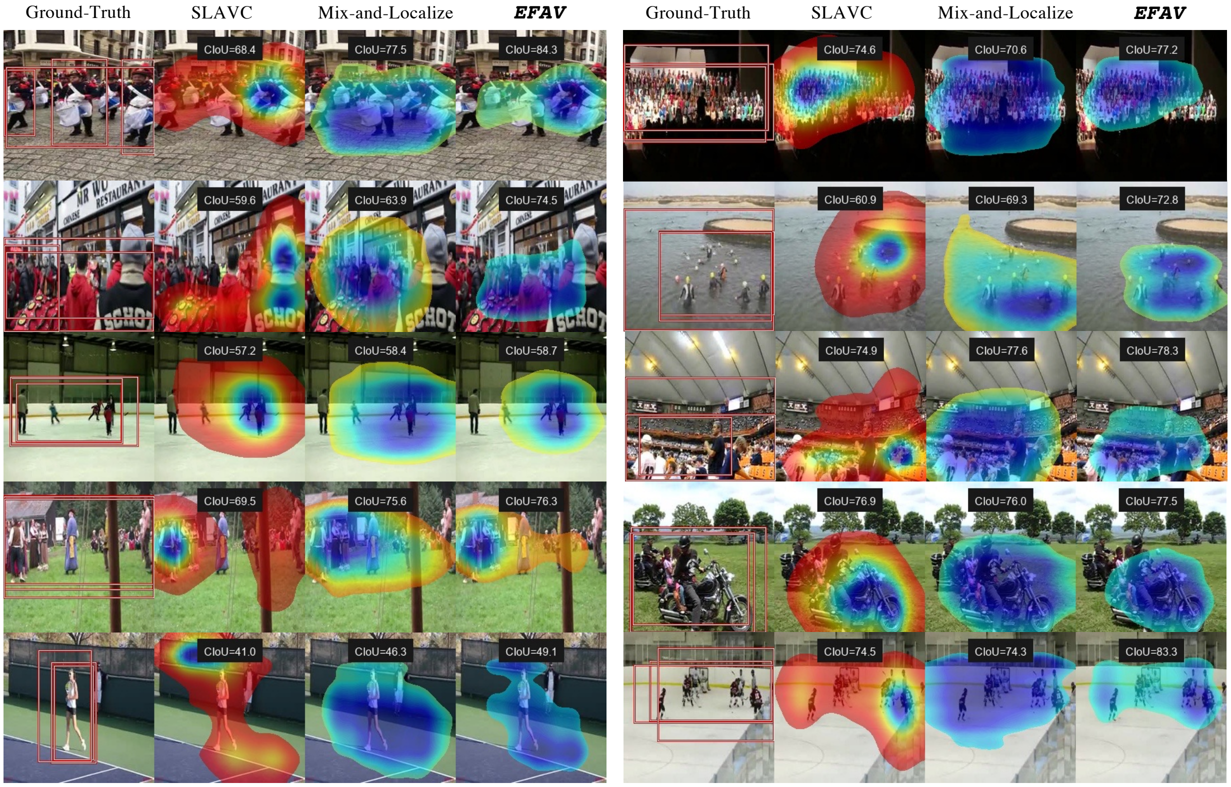

Figure 1. Qualitative visualization of sound source localization. The proposed DeepAVFusion produces more accurate and high-quality localization maps for each source.

and binary target masks, which identify the time-frequency bins where the target source is the most dominant component in the mixture. The model is trained for 20 epochs using the Adam optimizer [2] with a learning rate of $1e-4$ and a batch size of 128.

B. Demonstrations of model capabilities

B.1. Sound Source Localization and Segmentation

In order to qualitatively evaluate the localization and segmentation masks, we compare the proposed DeepAVFusion with SLAVC [3] and Mix-and-Localize [1] on sound source localization and segmentation in Figure 1 and Figure 2. We can observe that the quality of localization maps and masks generated by our framework are indeed superior to prior state-of-the-art methods. For example, the performance gains of DeepAVFusion over Mix-and-Localize [1] (a strong multi-task approach with both localization and separation) can be easily seen in these demonstrations. These visualizations showcase the effectiveness of the proposed DeepAVFusion in sound source localization and segmentation.

B.2. Sound Source Separation

In Figure 3, we qualitatively compare the proposed DeepAVFusion with audio-visual separation baselines, Sound-of-Pixels [7], CCoL [6], and OneAVM [4] in terms of reconstructed source spectrograms. Once again, we can observe the higher quality of spectrograms generated by our method, even when compared to OneAVM [4], the state-of-the-art model which jointly optimized for recognition, localization, and source separation. These visualizations further showcase the superiority of the proposed DeepAVFusion in sound source separation.

References

[1] Xixi Hu, Ziyang Chen, and Andrew Owens. Mix and localize: Localizing sound sources in mixtures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10483–10492, 2022. 2

[2] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1, 2

[3] 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

[4] Shentong Mo and Pedro Morgado. A unified audio-visual learning framework for localization, separation, and recognition. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2023. 1, 2

[5] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual

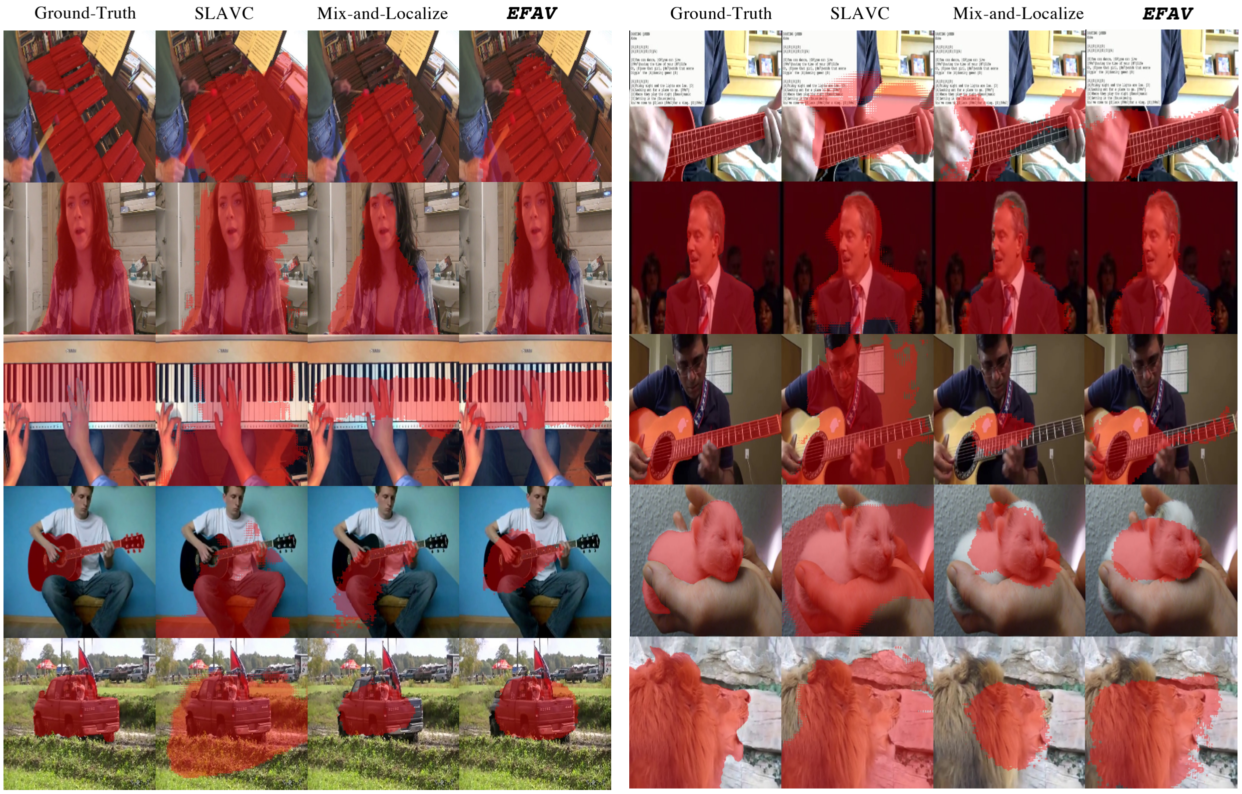

Figure 2. Qualitative visualization of sound source segmentation. The proposed DEEPAVFUSION produces more accurate and high-quality segmentation masks for each source.

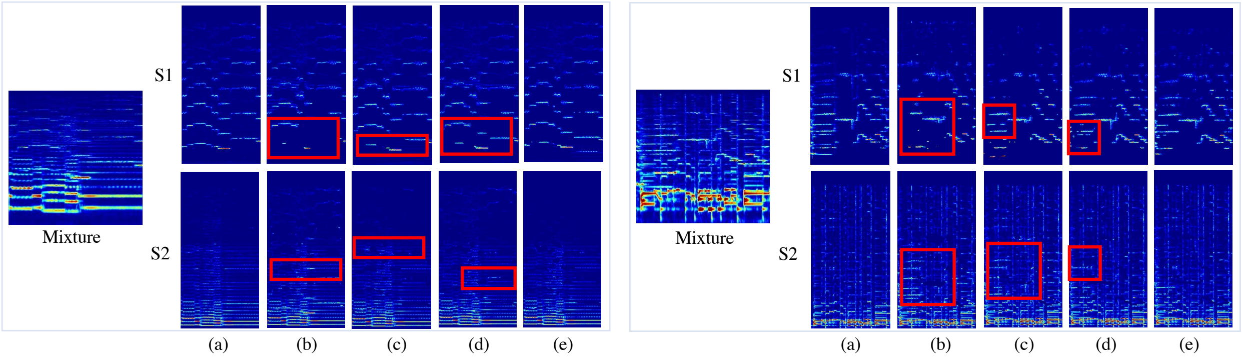

Figure 3. Qualitative visualization of sound source separation. (a) Ground-Truth; (b) Sound-of-Pixels; (c) CCoL; (d) OneAVM; (e) DEEPAVFUSION (ours). The proposed DEEPAVFUSION separates each source more accurately.

scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4358–4366, 2018. 1

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

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

[8] Jinxing Zhou, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, and Yiran Zhong. Audio-visual segmentation. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2022. 1