Learning to Visually Localize Sound Sources from Mixtures without Prior Source Knowledge
–Supplementary Material–

This manuscript provides additional implementation details and additional results of the proposed method. In Section 1, we elaborate on the additional implementation details of our method. Section 2 presents additional experimental results to show the effectiveness of the Iterative Object Identification (IOI) module and object similarity-aware clustering (OSC) loss. Moreover, Section 3 shows additional visualization results. Note that [PXX] indicates the reference in the main paper.

1. Additional Implementation Details

We utilize the ResNet-18 [P13] for the audio encoder, as mentioned in the main paper. Since the audio spectrogram has only one channel, we modify the first convolution layer of the encoder to have an input channel of 1 and an output channel of 64, utilizing a kernel size of 7, stride of 2, and padding of 3. Additionally, we employ the Adam optimizer, setting the parameters (β_1, β_2) to (0.9, 0.999), which are the standard values for Adam. For the hyperparameter θ and ω, mentioned in Section 3.2, adopt the values 0.65 and 0.03, respectively, following [P8].

2. Additional Experiments

Training Time per Epoch in Training Phase. Since our approach adopts an iterative method, we explored how training duration varies across epochs, as illustrated in Figure 1. Initially, in the first epoch, the processing time is high at 1,594 seconds per epoch. However, with more epochs, this time significantly drops and stabilizes around 500 seconds per epoch. This trend suggests that our method becomes more computationally efficient over time by reducing unnecessary iterations, focusing on the effective steps for localizing sound-making objects.

Comparison of our method with baseline (single localization applied after separation). We present an additional experiment to validate the robustness of our approach for localizing sound sources from mixtures by comparing it with a baseline method which is first to perform audio source separation on the mixture and then apply single sound source localization to each segregated audio element. On MUSIC-Duet, our method is superior to the baseline (widely used audio separation model [1] followed by single SSL), showing CAP (22.4→52.1), PIAP (44.8→72.5), CIoU@0.3 (29.8→38.6), and AUC (23.6→30.1).

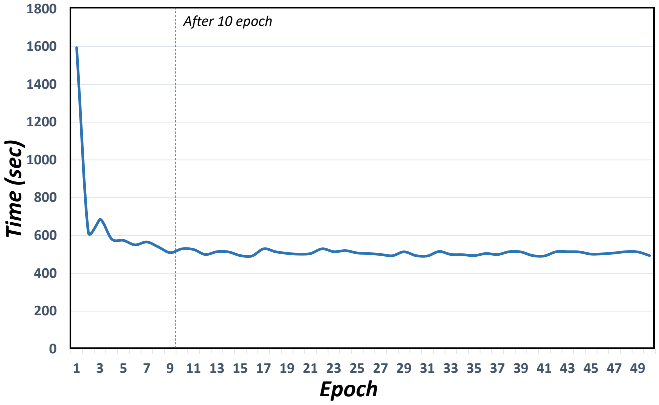

Figure 1. Visualization of training time per epoch of our model. After epoch 10 (red line), the training time converges to about 500 seconds/epoch.

Impact of audio component. We conducted the visual-only experiment on the MUSIC-Duet data set to investigate the significance of the audio component. Adding an audio component enhances performance across all metrics: CAP (20.5→52.1), PIAP (31.4→72.5), CIoU@0.3 (26.1→38.6), and AUC (21.2→30.1).

3. Additional Visualization Results

Visualization Results on VGGSound-Duet, Trio and Mixed Dataset. We present additional visualization results of our method to demonstrate its efficacy in differentiating objects in scenarios with various source mixtures utilizing a VGGSound-Duet, Trio, and VGGSound-Mixed test set. VGGSound-Trio test set is comprised mixture of three sound sources from VGGSound-Single [P7], as guided by [P26], and VGGSound-Mixed test set. Our IOI module is adept at repeatedly detecting and distinguishing sound-making objects within audio-visual scenes, resulting in highly accurate and detailed localization maps. These visualizations, as depicted in Figures 2, 3 and 4, demonstrate the accuracy of our model in individual object localization and total map estimation, reflecting a deep understanding of the complex audio-visual landscape.

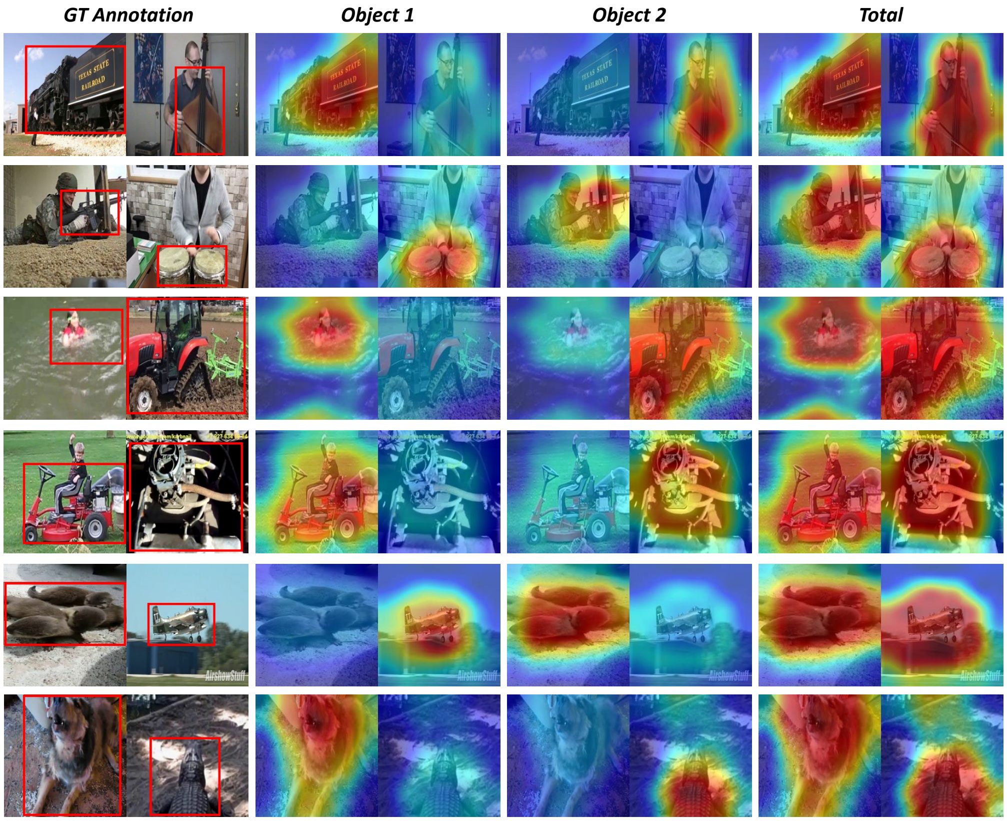

Figure 2. Additional visualization results for VGGSound-Duet test set (two objects). 'Object k' is identified by our model without any prior knowledge.

Visualization Results without OSC loss. Based on the ablation study on the effect of the proposed losses within our manuscript, we present visualization results for scenarios with and without the OSC (Object Similarity-aware Clustering) loss. This comparison is based on two samples used in Figure 5 of the manuscript. In Figure 5, the results demonstrate that the use of OSC loss leads to better performance in separating objects in mixtures. Consequently, it can be observed that the ability of our model to distinguish between objects is enhanced through the incorporation of the OSC loss.

Video Demo. We provide video materials that offer a more in-depth explanation of our method for localizing sound-making objects in complex environments. These videos demonstrate the real-time applicability and robustness of our approach under various conditions. We provide results of our method with some examples from the VGGSound-Duet dataset. Please see video in our official repository.

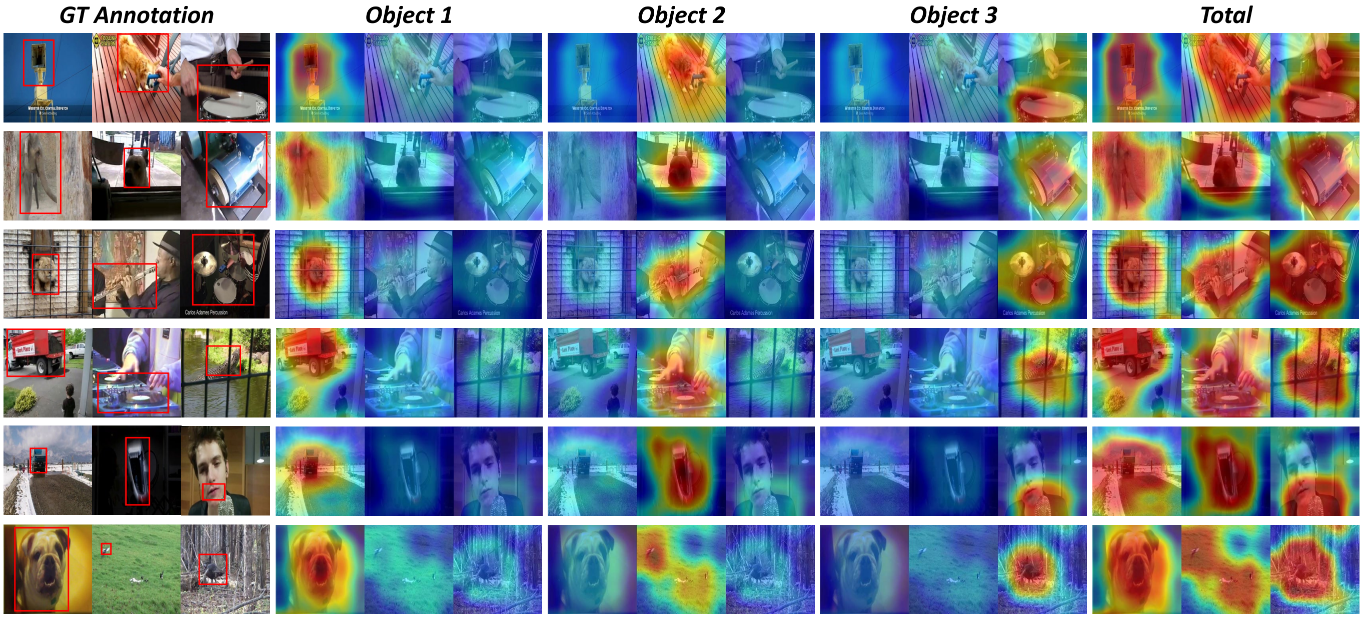

Figure 3. Additional visualization results for VGGSound-Trio test set (three objects). ‘Object k’ is identified by our model without any prior knowledge.

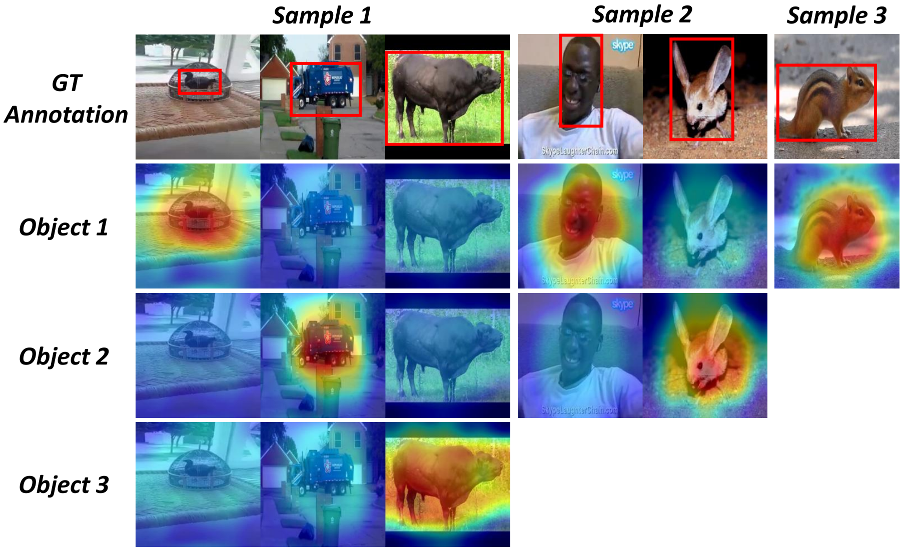

Figure 4. Additional visualization results for VGGSound-Mixed test set (mixed objects). ‘Object k’ is identified by our model without any prior knowledge.

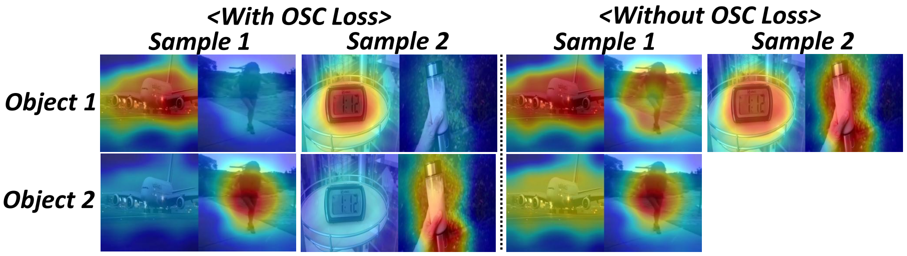

Figure 5. Additional visualization results with/without OSC loss. ‘Object k’ is identified by our model without any prior knowledge.

References

[1] Subakan et al. Attention is all you need in speech separation. In *ICASSP*, 2021.

[P7] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP*, 2020.

[P8] Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Localizing visual sounds the hard way. In *CVPR*, 2021.

[P13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[P26] Shentong Mo and Yapeng Tian. Audio-visual grouping network for sound localization from mixtures. In *CVPR*, 2023.