

Robust Audio-Visual Segmentation via Audio-Guided Visual Convergent Alignment

Chen Liu^{1,4}, Peike Li³, Liying Yang⁵, Dadong Wang⁴, Lincheng Li², Xin Yu^{1*}

¹ The University of Queensland, ² NetEase Fuxi AI Lab, ³ Matrix Verse AI,

⁴ CSIRO Data61, ⁵ Macau University of Science and Technology

yenianliu36@gmail.com, xin.yu@uq.edu.au

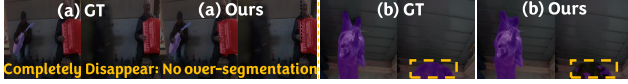


Figure 1. Failure Case Analysis. (a) illustrates cases where audible objects have completely disappeared, while (b) presents the model’s performance when audible objects are partially occluded.

Table 1. Comparison results with metrics in TPAVI [14].

Methods	AVS-S4		AVS-MS3		AVSS		VPO-SS		VPO-MS		VPO-MSMI	
	\mathcal{J}	\mathcal{F}_m	\mathcal{J}	\mathcal{F}_m	\mathcal{J}	\mathcal{F}_m	\mathcal{J}	\mathcal{F}_m	\mathcal{J}	\mathcal{F}_m	\mathcal{J}	\mathcal{F}_m
CAVP	67.5	75.9	45.2	57.1	32.2	37.1	36.8	40.0	35.1	39.5	37.8	39.3
SESI	83.5	91.2	60.3	71.3	41.3	46.9	42.3	48.8	42.8	49.4	42.5	48.5
AVSBias	82.9	92.8	66.1	79.2	41.9	47.6	42.2	46.7	41.5	44.2	42.4	47.2
AVSSone	83.2	91.3	67.3	77.6	48.5	53.2	41.5	47.3	42.1	47.3	41.9	46.2
Ours	85.7	93.5	69.2	78.9	50.7	55.3	46.3	49.1	44.1	51.6	44.8	50.7

0.1. Failure Case Analysis.

Fig. 1 (a) shows the model does not segment other silent regions when the audible object completely disappears. Fig. 1(b) indicates that when an object is partially visible, the significant reduction in visual information weakens the AV alignment, thus degrading segmentation quality.

Table 2. More comparisons with recent methods.

Methods	AVS-Object-S4			AVS-Object-MS3			AVSS		
	\mathcal{J}	\mathcal{F}_β	\mathcal{J}	\mathcal{F}_β	\mathcal{J}	\mathcal{F}_β	\mathcal{J}	\mathcal{F}_β	\mathcal{J}
[13] [TCSVT24]	87.1	83.3	90.8	72.1	67.3	77.0	-	-	-
[2] [ICPR24]	88.1	84.5	91.6	70.4	64.2	76.6	39.7	42.4	37.0
[8] [WACV24]	85.1	81.5	88.6	66.2	63.1	69.1	-	-	-
Ours	89.6	85.7	93.5	74.1	69.2	78.9	53.0	50.7	55.3

0.2. Module Effectiveness Analysis.

We visualize the segmentation results from the ablation of CST and UE. Fig. 2 shows the model with CST accurately segments sounding objects. Meanwhile, the model with UE avoids over-segmentation when the sound state changes.

0.3. More Method Comparisons.

Table 2 shows that our method consistently outperforms the existing methods [2, 8, 13] on the AVS datasets with metrics in TPAVI.

*Corresponding author.



Figure 2. Visual results of the ablation study.

1. More Implementation Details

1.1. Implementation Details.

Our framework is trained on eight NVIDIA V100 GPUs in parallel mode. In particular, we train 80 epochs on the AVSS dataset and 40 epochs on the VPO dataset. Following the standard setup in [4, 6, 15], we employ a crop size of 224×224 pixels for all visual frames across experiments. For data augmentation, images are randomly flipped horizontally, adjusted with color jitter, and scaled within a ratio range of 0.5 to 2.0. For network optimization, we utilize AdamW [9] with an initial learning rate of 1×10^{-4} , epsilon set to 1×10^{-8} , and betas set to [0.9, 0.999]. We apply a warmup strategy with 2 warmup epochs and a warmup learning rate of 4×10^{-6} . For initialization, we load pre-trained weights from ImageNet [3] for the visual model and from AudioSet [5] for the audio model, as in [7, 10, 11, 15].

1.2. Network Configuration.

Our framework is end-to-end trainable, with all components parameterized by neural networks. We employ MIT-B5 [12] as our visual backbone for visual feature extraction, while the audio encoder is based on HT-SAT [1], which is frozen during the training process. The decoder comprises MLP layers for fused multi-scale feature maps, convolutional layers for spatial and temporal uncertainty estimation, and a Multi-Head Attention layer to capture temporal dynamics across the $T \times D$ dimension. For the audio-guided modality alignment process, the thresholds σ_a for determining positive and negative samples is set to 0.5, and the temperature parameter τ is set to 0.1.

References

- [1] Ke Chen, Xingjian Du, Bilei Zhu, Zejun Ma, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Hts-at: A hierarchical token-semantic audio transformer for sound classification and detection. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 646–650. IEEE, 2022. [1](#)
- [2] Tianxiang Chen, Zhentao Tan, Tao Gong, Qi Chu, Yue Wu, Bin Liu, Nenghai Yu, Le Lu, and Jieping Ye. Bootstrapping audio-visual video segmentation by strengthening audio cues. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. [1](#)
- [3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. [1](#)
- [4] Shengyi Gao, Zhe Chen, Guo Chen, Wenhai Wang, and Tong Lu. Avsegformer: Audio-visual segmentation with transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12155–12163, 2024. [1](#)
- [5] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017. [1](#)
- [6] Xiang Li, Jinglu Wang, Xiaohao Xu, Xiulian Peng, Rita Singh, Yan Lu, and Bhiksha Raj. Qdformer: Towards robust audiovisual segmentation in complex environments with quantization-based semantic decomposition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3402–3413, 2024. [1](#)
- [7] Zhangbin Li, Dan Guo, Jinxing Zhou, Jing Zhang, and Meng Wang. Object-aware adaptive-positivity learning for audio-visual question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3306–3314, 2024. [1](#)
- [8] Jinxing Liu, Yu Wang, Chen Ju, Chaofan Ma, Ya Zhang, and Weidi Xie. Annotation-free audio-visual segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 5604–5614, 2024. [1](#)
- [9] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. [1](#)
- [10] Juncheng Ma, Peiwen Sun, Yaoting Wang, and Di Hu. Stepping stones: A progressive training strategy for audio-visual semantic segmentation. *arXiv preprint arXiv:2407.11820*, 2024. [1](#)
- [11] Peiwen Sun, Honggang Zhang, and Di Hu. Unveiling and mitigating bias in audio visual segmentation. *arXiv preprint arXiv:2407.16638*, 2024. [1](#)
- [12] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34: 12077–12090, 2021. [1](#)
- [13] Yinhao Zhang, Tianyang Xu, Xiao-Jun Wu, Shaochuan Zhao, and Josef Kittler. Multi-frequency fine-grained matching for audio-visual segmentation. In *International Conference on Pattern Recognition*, pages 36–50. Springer, 2024. [1](#)
- [14] 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 *European Conference on Computer Vision*, pages 386–403. Springer, 2022. [1](#)
- [15] Jinxing Zhou, Xuyang Shen, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, et al. Audio-visual segmentation with semantics. *International Journal of Computer Vision*, pages 1–21, 2024. [1](#)