Supp: Implicit Counterfactual Learning for Audio-Visual Segmentation

Mingfeng Zha[1], Tianyu Li[1], Guoqing Wang[1], Peng Wang[1], Yangyang Wu[2], Yang Yang[1], Heng Tao Shen[3]
[1]University of Electronic Science and Technology of China, [2]Zhejiang University, [3]Tongji University

Table 1. Hyperparameter settings.

Hyperparameter	Value
k^t	15
λ_{z}	0.8
λ_{ortho}	1.0
λ_{bce}	1.0
λ_{dice}	1.0
λ_{focal}	1.0
λ_{cf}	0.2
λ_{CDCL}	0.2
$\lambda_{\text{v}\leftrightarrow\text{a}}$	0.4
$\lambda_{\text{a}\leftrightarrow\text{l}}$	0.3
$\lambda_{\text{v}\leftrightarrow\text{l}}$	0.3

We provide segmentation results on three datasets, which can be found in the `segmentation_results.zip`.

1. Evaluation Metrics

Follow [1, 3], we adop $\mathcal{J}$ to measure segmentation quality, $\mathcal{F}$ to quantify precision and recall results, and false detection rate (FDR) to illustrate misclassification of categories.

2. Segmentation Results

In Figure 1, considering the complexity of visual and audio information, we select four cases from the AVSS dataset (from top to bottom). Case 1: AVSBench incorrectly classifies pixels and is incomplete, while AVSegformer exhibits discontinuous segmentation and lacks sufficient temporal modeling. Case 2: Contrasting methods struggle to determine the target outline (person) and may even miss detections. Case 3: AVSBench lacks global dependency perception within frames and misclassifies instruments into other categories, a deficiency partly alleviated by AVSegformer. Case 4: Contrasting methods fail to effectively perceive rapidly moving targets. In Table **??**, we evaluate the performance of each category (50/70) under different metrics, providing fine-grained comparison.

3. Implementation Details

1. For the S4 and M3 datasets, the time window is set to $\lceil \frac{T}{2} \rceil$; for the AVSS dataset, it is set to $\lceil \frac{T}{4} \rceil$.
2. We empirically set k^t to 30. Generally, a larger k^t increases redundancy, while a smaller value may result in capturing overly coarse semantic granularity.
3. The visual features $\mathbf{F}^v$ and audio features $\mathbf{A}$ used for the CDCL are processed by the decoder (after interaction), rather than being directly extracted by the encoder.
4. In the baseline, we leverage videos from MIT, employing frame-level channel-wise cross-attention and self-attention to enhance the spatiotemporal context modeling of visual features. We utilize bidirectional queries to prevent the dispersion of features that may occur when audio is used solely as the query.
5. For the signal-to-noise ratio setting, we follow [2]. For frame mixing, we select 1/4 of the M3 data and randomly swap it with other video frames.
6. In the setting of the ablation contrast strategy, we apply masked average pooling to generate visual prototypes. And We use the mean to generate audio prototypes.

References

[1] Yuanhong Chen, Yuyuan Liu, Hu Wang, Fengbei Liu, Chong Wang, and Gustavo Carneiro. A closer look at audio-visual semantic segmentation. *arXiv e-prints*, pages arXiv–2304, 2023. 1

[2] 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

[3] 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

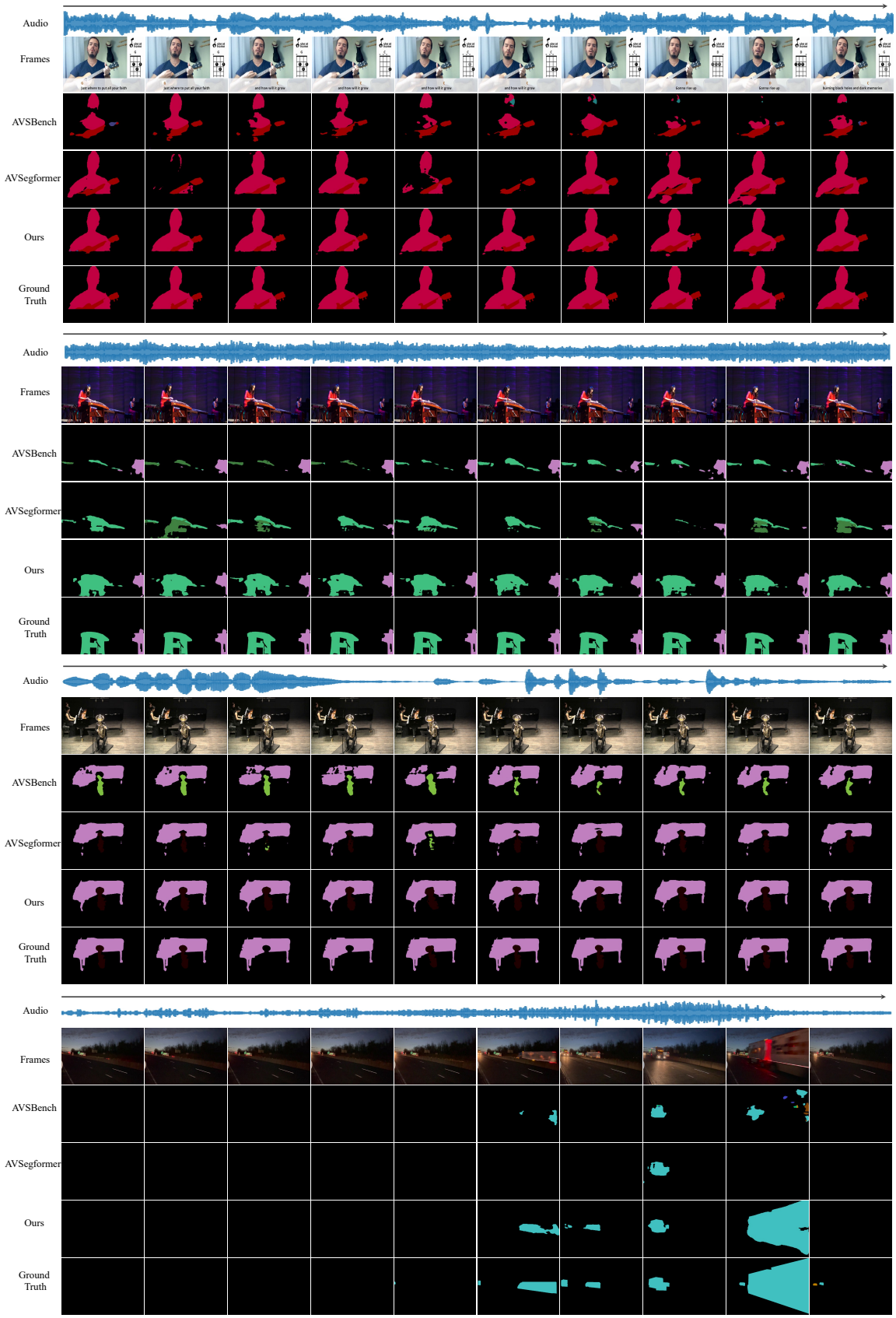

Figure 1. Qualitative comparison on the AVSS benckmark.

Figure 2. Comparison of t-SNE visualization with and without the CDCL.