LightAVSeg: Lightweight Audio-Visual Segmentation

Qing Zhong[1], Guodong Ding[2*], Lingqiao Liu[3], Zaiwen Feng[1], Lin Yuanbo Wu[4 5], Angela Yao[2]

[1] Huazhong Agricultural University, [2] National University of Singapore, [3] Adelaide University, [4] University of Warwick, [5] Zhejiang Yuexiu University

1. Task & Motivation

Audio-Visual Segmentation

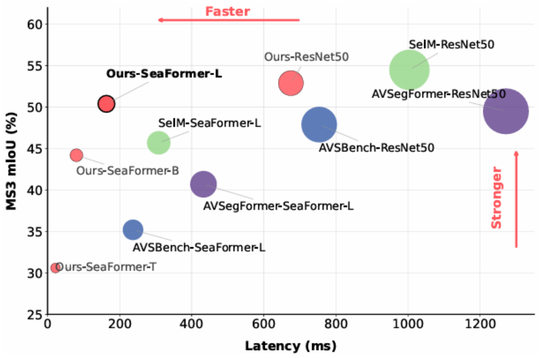
Pixel-level localization of sound-emitting objects in videos.

Lightweight AVS

Enable accurate AVS on resource-constrained devices, e.g., mobile phones, VR headsets, and on-device video editing.

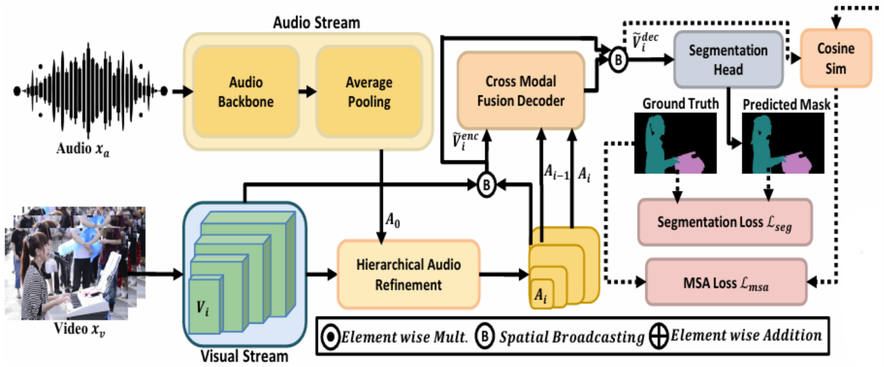
Challenges

1. Existing AVS methods rely on dense cross-modal attention.
2. The interaction cost grows quadratically with spatial tokens.
3. Backbone replacement alone does not solve the efficiency bottleneck.
4. Lightweight models may learn spurious audio-visual correlations.



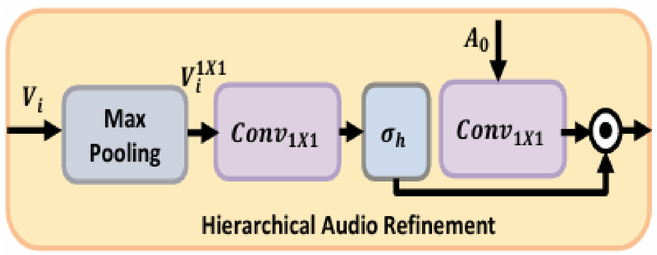
2. Approach

A lightweight AVS framework that decouples audio-visual interaction into semantic filtering and spatial grounding.



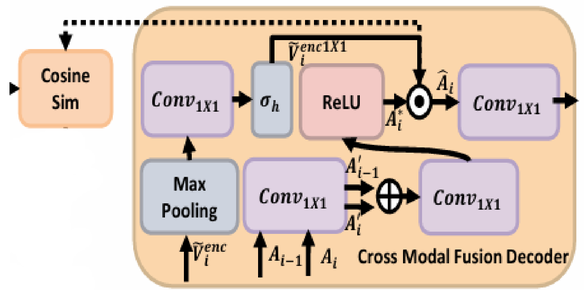
Reciprocal Audio-Visual Encoder

1. Refine the global audio state using visual context.
2. Modulate visual features with refined audio cues.
3. Replace dense attention with channel-wise interaction.
4. Reduce interaction complexity from $O(N^2)$ to $O(N)$..



Cross-Modal Fusion Decoder

1. Progressively recover spatial resolution.
2. Maintain audio-guided semantic consistency.
3. Inject refined audio cues into visual features.
4. Generate the final sounding-object mask.



Multi-Scale Audio-Visual Alignment Loss

1. Align audio cues with valid visual regions.
2. Guide the decoder to focus on sounding objects.
3. Training-only supervision with zero inference overhead.

$$\mathcal{L}_{\text{msa}} = \frac{1}{S}\sum_{i=1}^{S} \text{BCE}(\hat{s}_i,\ M)$$

Total Loss

Lseg supervises the final segmentation mask.
Lmsa aligns audio cues with sounding visual regions.
Training-only loss with zero inference overhead.

$$\mathcal{L}_{\text{seg}} = \mathcal{L}_{\text{dice}} + \mathcal{L}_{\text{bce}}$$

$$\mathcal{L} = \mathcal{L}_{\text{seg}} + \lambda \mathcal{L}_{\text{msa}}$$

3. Results

Performance

1. New lightweight SOTA on S4, MS3, and AVSS.
2. Achieves 75.6 MJ on S4 and 50.4 MJ on MS3.
3. Surpasses AVSegFormer-R50 on MS3 with only 20.5M parameters.
4. Runs at 163.4 ms on Snapdragon 8 Elite.

Method	Backbone		GPU	Mobile	S4		MS3		Params
	Visual	Audio	ms ↓	ms ↓	M_J ↑	M_F ↑	M_J ↑	M_F ↑	
AVSBench (Zhou et al., 2022)	R50	VGGish	21.2	753.5	72.8	84.8	47.9	57.8	91.4M
AVSegFormer (Gao et al., 2024)	R50	VGGish	29.0	1271.4	76.5	85.9	49.5	62.8	151.1M
SelM (Li et al., 2024)	R50	VGGish	25.3	1003.8	76.6	86.2	54.5	65.6	117.6M
AVSBench (Zhou et al., 2022)	Sea	MNetV2	19.5	237.1	47.9	64.5	35.2	44.1	30.2M
AVSegFormer (Gao et al., 2024)	Sea	MNetV2	22.2	432.6	53.8	71.4	40.7	50.7	51.0M
SelM (Li et al., 2024)	Sea	MNetV2	18.7	308.6	59.1	77.4	45.7	57.6	39.5M
Ours	Sea	MNetV2	**15.9**	**163.4**	**75.6**	**86.2**	**50.4**	**62.6**	**20.5M**

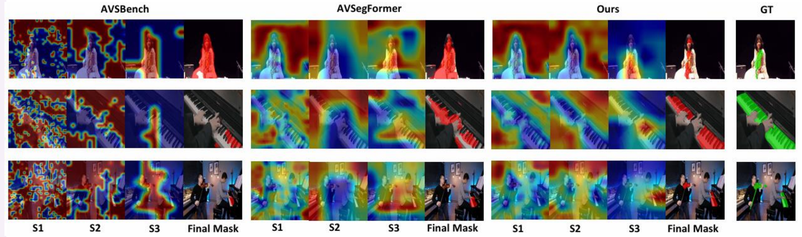
Ablation Study

1. Baseline static fusion: 44.4 MJ / 56.8 MF.
2. Audio-guided visual enhancement improves to 46.8 MJ.
3. Hierarchical audio refinement improves to 49.3 MJ.
4. Full model with CMFD and Lmsa achieves 50.4 MJ / 62.6 MF.

AGVE	CMFD	HAR	$\mathcal{L}_{\text{msa}}$	M_J ↑	M_F ↑
				44.4	56.8
✓				46.8	58.1
✓	✓			48.3	59.1
✓		✓		49.3	61.2
✓	✓	✓	✓	50.4	62.6

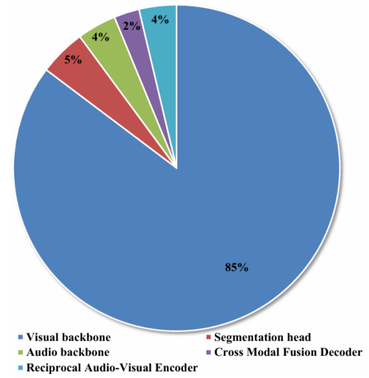
Qualitative Result

1. AVSBench is distracted by background noise.
2. AVSegFormer focuses on targets but lacks clear boundaries.
3. Ours shows coarse-to-fine refinement.
4. Ours progressively suppresses background and aligns with GT.



Latency Analysis

1. Visual backbone dominates inference time.
2. RAVE and CMFD take only 6.1 ms and 4.1 ms.
3. Cross-modal interaction introduces negligible overhead.



4. Takeaways

Efficient Interaction

Lightweight AVS requires efficient cross-modal interaction,not only a smaller backbone.

What-Where Decoupling

Audio tells the model what is sounding,while visual features determine where to segment.

Training-only Alignment

Lmsa improves audio-visual consistency during trainingwith zero inference overhead.

Practical Mobile AVS

LightAVSeg achieves strong accuracy-efficiency trade-off,enabling high-performance AVS on mobile devices.