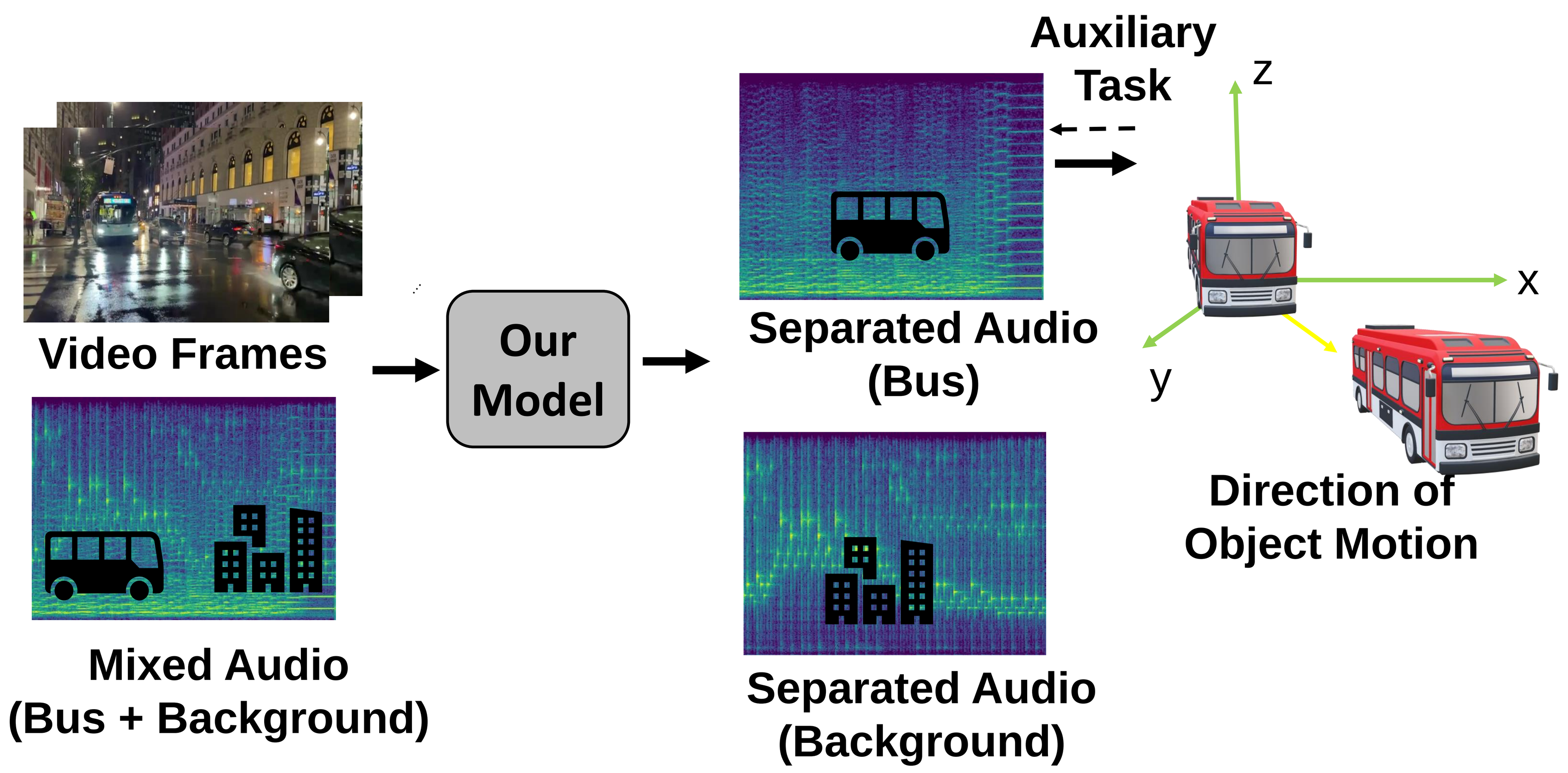
Learning Audio-Visual Dynamics Using Scene Graphs for Audio Source Separation

Moitreya Chatterjee*[1,2], Narendra Ahuja[1], and Anoop Cherian* [2]

University of Illinois, Urbana-Champaign (UIUC)[1], Mitsubishi Electric Research Laboratories (MERL)[2]

https://sites.google.com/site/metrosmiles/research/research-projects/asmp

Problem Statement



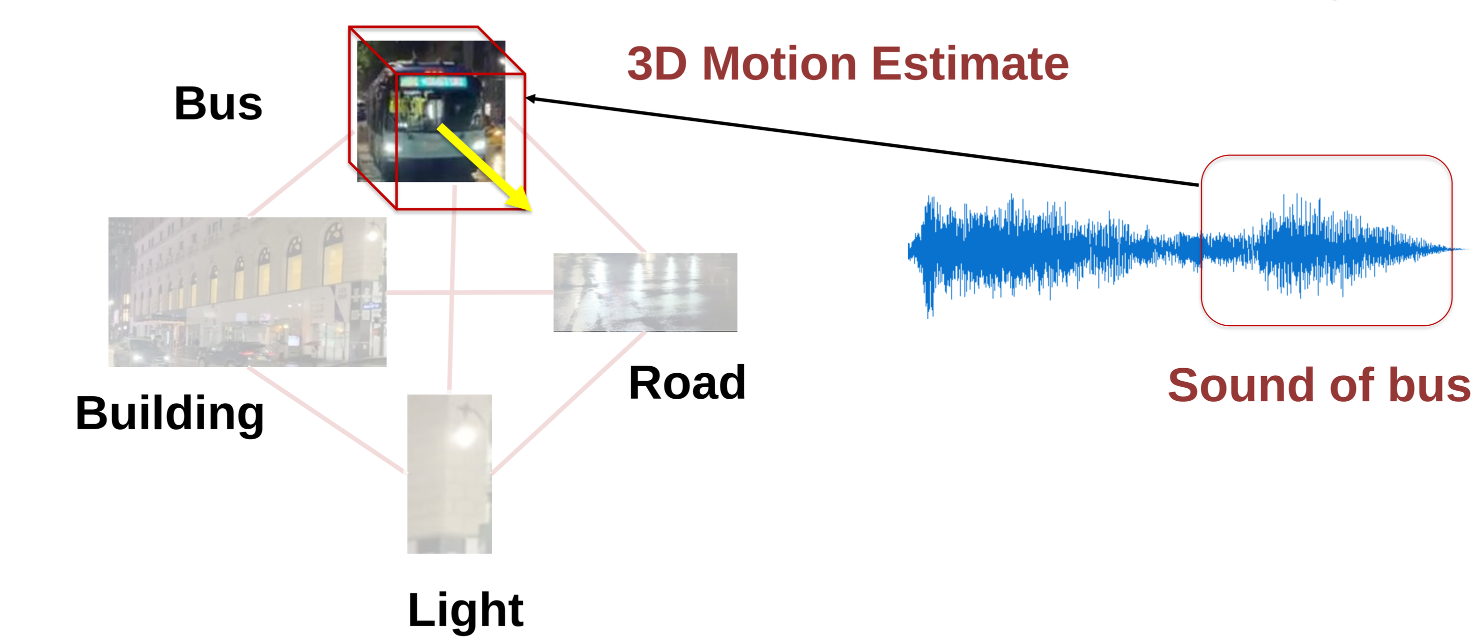
- We study the task of **visually-guided audio source separation**, i.e., given an audio mixture of multiple sound sources, the task is to separate it into its constituents using the available **visual information**.
- We leverage pseudo-3D scene geometry information encoded via scene-graphs and directionality of the object's motion to accomplish this.

Prior Work

- **Gao et al. (ICCV'19)**: Uses visual information but neither the visual context nor motion is leveraged for this task.
- **Zhao et al. (ICCV'19)**: They incorporate object motion, but the 3D nature of the scene is not exploited.
- **AVSGS (ICCV'21)**: Here the visual context of the object is incorporated into the visual representation, but the 3D geometry is not.

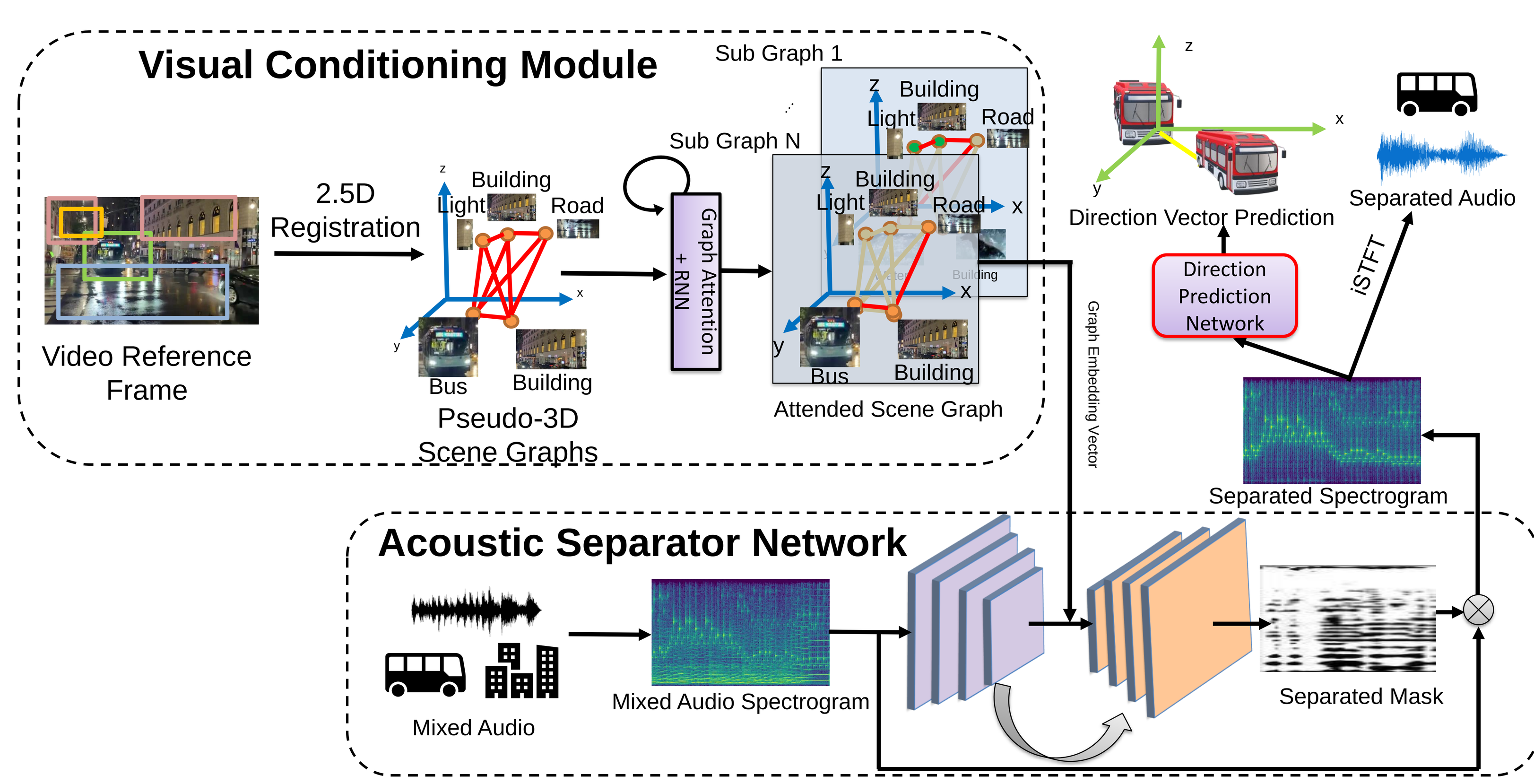
Audio Separation and Motion Prediction

Audio-Visual Scene Graphs



- We present a <u>2.5D geometry aware scene-graph</u> based approach for the task of <u>visually guided audio source separation</u> called **Audio Separator and Motion Predictor (ASMP)**.
- We predict the <u>direction of motion</u> of the sound source, aided by appropriate visual context, to derive <u>additional supervision</u> for training our model.

Model Architecture and Losses



- **Orthogonality:** $\mathcal{L}_{\text{ortho}}(Y) = \sum_{i,j\in\{1,2,\dots,N\}, i\neq j} (y_i^\top y_j)^2$
- **Consistency:** $\mathcal{L}_{\text{cons}} = \sum_{u=1,2} \min_{\sigma^u\in\mathcal{S}_{N_u+1}} -\sum_{i=1}^{N_u+1}\sum_{c=1}^{K} \mathbb{1}^u_{i,\sigma^u(c)} \log p^u_{i,c}$
- **Cyclic:** $\mathcal{L}_{\text{cyc}} = \sum_{u=1,2} \left\| \sum_{i=1}^{N_u+1} \hat{\mathbf{M}}_i^u - \mathbf{M}^u_{\text{ibm}} \right\|_1$
- **Direction Pred:** $\mathcal{L}_{\text{dirpred}} = \sum_{u=1,2}\sum_{w=1}^{W} \min_{\sigma^u\in\mathcal{S}_{N_u+1}} -\sum_{i=1}^{N_u+1}\sum_{c=1}^{D_k} \mathbb{1}^{u,w}_{i,\sigma^u(c)} \log q^{u,w}_{i,c}$

Experimental Analysis

Quantitative Study

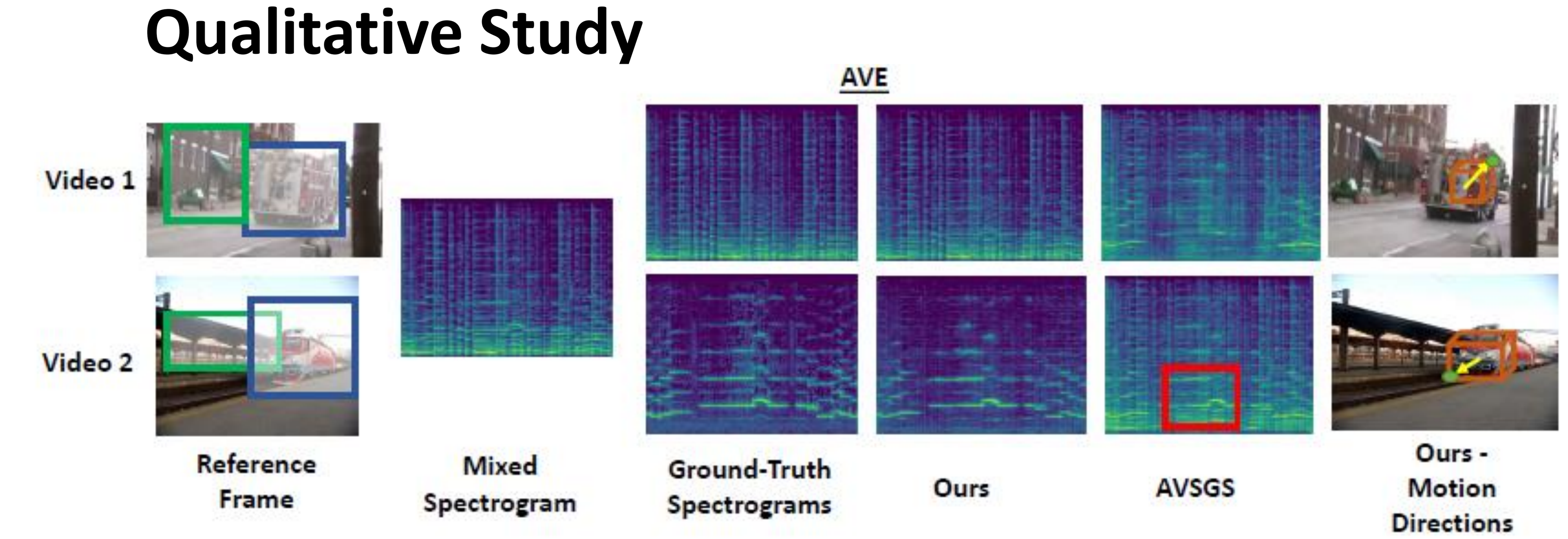
Table 1: SDR, SIR, and SAR results on the ASIW and AVE test sets. [Key: **Best**, second-best results.]

Approach	ASIW			AVE		
	SDR ↑	SIR ↑	SAR ↑	SDR ↑	SIR ↑	SAR ↑
Sound of Motion (SofM) [55]	6.7	9.4	11.1	4.1	9.2	7.6
Cyclic Co-Learn [46]	7.0	13.4	12.4	4.2	9.7	8.4
Co-Separation [13]	6.6	12.9	12.6	3.9	9.3	7.8
AVSGS [8]	8.8	14.1	13.0	5.8	10.4	8.2
ASMP (only 2.5D graph)	**9.0**	**14.3**	**13.7**	**6.5**	**12.4**	**8.9**
ASMP (2.5D graph + motion)	**9.6**	**14.5**	**14.1**	**7.2**	**13.3**	**9.4**

Table 2: Direction Prediction results on the ASIW and AVE on test splits.

Direction Prediction	ASIW		AVE	
	10-class (%)↑	28-class (%)↑	10-class (%)↑	28-class (%)↑
Majority Vote	27.3	25.4	29.2	24.3
Sound of Motion (SofM) [55]	29.6	27.0	31.2	30.6
Cyclic Co-Learn [46]	34.8	32.3	30.7	29.2
Co-Separation [13]	32.2	31.7	30.2	28.0
AVSGS [8]	39.2	38.7	**38.9**	34.7
ASMP (Ours)	**42.5**	**41.3**	38.5	**36.8**

Qualitative Study



Conclusions

- We explore the efficacy of geometry-aware visual representation and motion cues for the task of visually guided audio source separation.
- We propose a novel 2.5D scene-graph representation (ASMP) towards this end and train it using weakly-/self-supervised loses such as predicting the direction of motion.
- We achieve state-of-the-art results on two challenging audio-visual datasets.

Acknowledgements

MC initiated that work at UIUC and completed it at MERL. MC was partially supported, and NA was fully supported by ONR under grant N00014- 20-1-2444, and USDA National Institute of Food and Agriculture under grant 2020-67021-32799/1024178. AC was fully supported by MERL.