Listen to the Pixels

Sanjoy Chowdhury[1,2], Subhrajyoti Dasgupta[3], Sudip Das[3], Ujjwal Bhattacharya[3]

[1] International Institute of Information Technology, Hyderabad
[2] ShareChat, Bangalore
[3] Indian Statistical Institute, Kolkata

Paper #2351

Audio-visual Co-Segmentation

Frames showing a moving sound-producing object

Audio-visual Co-Segmentation

Localization of the sound-producing objects
&
Separation of the sound sources

Audio-visual Co-Segmentation

Applications

- Understanding which parts of the image are producing sound
- Independent volume control of different sound sources
- Removal of specific audio sources
- Independent audio adjustments of different sound sources
- Moving vehicle tracking [2] and others.

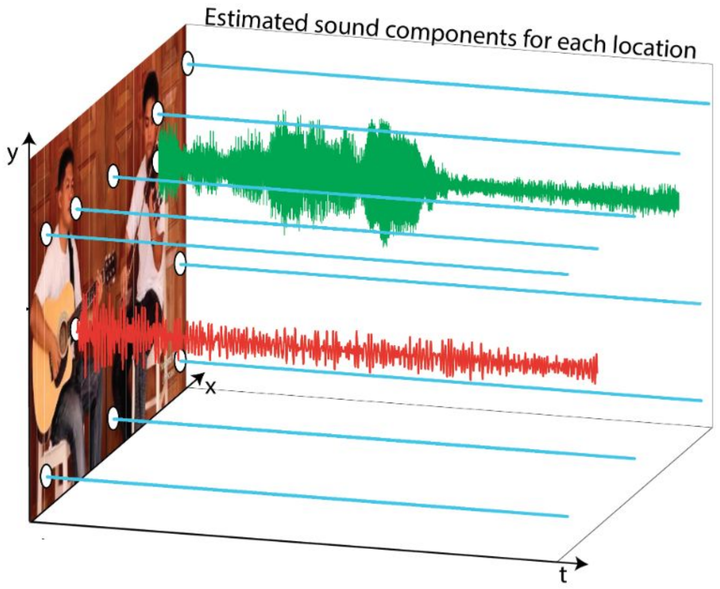


Fig. - Audio-visual Co-Segmentation [1]

[1] - Zhao, Hang, et al. "The sound of pixels." *Proceedings of the European conference on computer vision (ECCV)*. 2018.

[2] - Gan, Chuang, et al. "Self-supervised moving vehicle tracking with stereo sound." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019.

Challenges

Only one sound-producing object among many others

Multiple similar sound-producing objects

Sound-producing objects in-the-wild

One sound-producing object occluding the other(s)

Distant sound-producing objects

Lack of annotated data

Related Works

- An unsupervised learning algorithm for the separation of sound sources in one-channel music signals [1]
- A network that can localize the object that sounds in an image, given the audio signal [2]
- PixelPlayer - a system to locate image regions which produce sounds and separate the input sounds that represents the sound from each pixel [3]
- Audio-visual event localization by jointly taking both audio and visual features at each time segment as inputs [4]

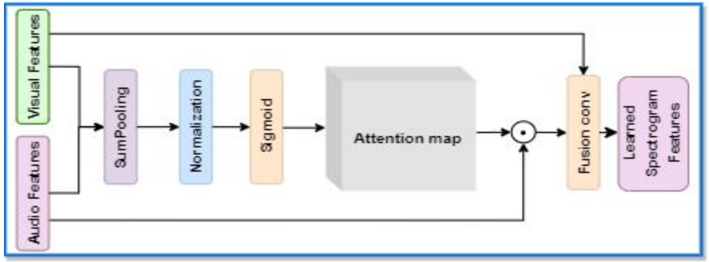
[1] Tuomas Virtanen, “Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1066–1074, 2007

[2] Arandjelovic, Relja, and Andrew Zisserman. "Objects that sound." *Proceedings of the European conference on computer vision (ECCV)*. 2018.

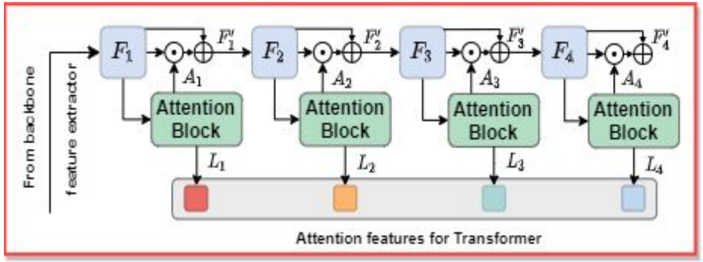
[3] Zhao, Hang, et al. "The sound of pixels." *Proceedings of the European conference on computer vision (ECCV)*. 2018.

[4] Lin, Yan-Bo, Yu-Jhe Li, and Yu-Chiang Frank Wang. "Dual-modality seq2seq network for audio-visual event localization." *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019.

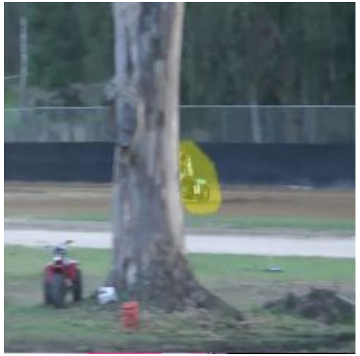
Our Contributions



Efficient blending of audio-visual information through LoGAn



A novel *Spatial Attention Block*



Partially occluded sound-source segmentation

Audio intensity cue guided segmentation of multiple sound-sources

Table 1: Performance comparison with respect to sound separation and semantic segmentation.

Method	SDR	SIR	Visual Segmentation Accuracy (%)
Audio feature only	5.28	9.43	59.68
Visual feature only	4.16	6.88	63.49
Zhao et al. [6]	1.03	6.37	45.90
PixelPlayer [5]	4.96	9.21	64.42
AViS-Net [ours]	**7.43**	**13.16**	**70.95**

Outperforming existing SOTA methods in joint audiovisual segmentation in unconstrained setting

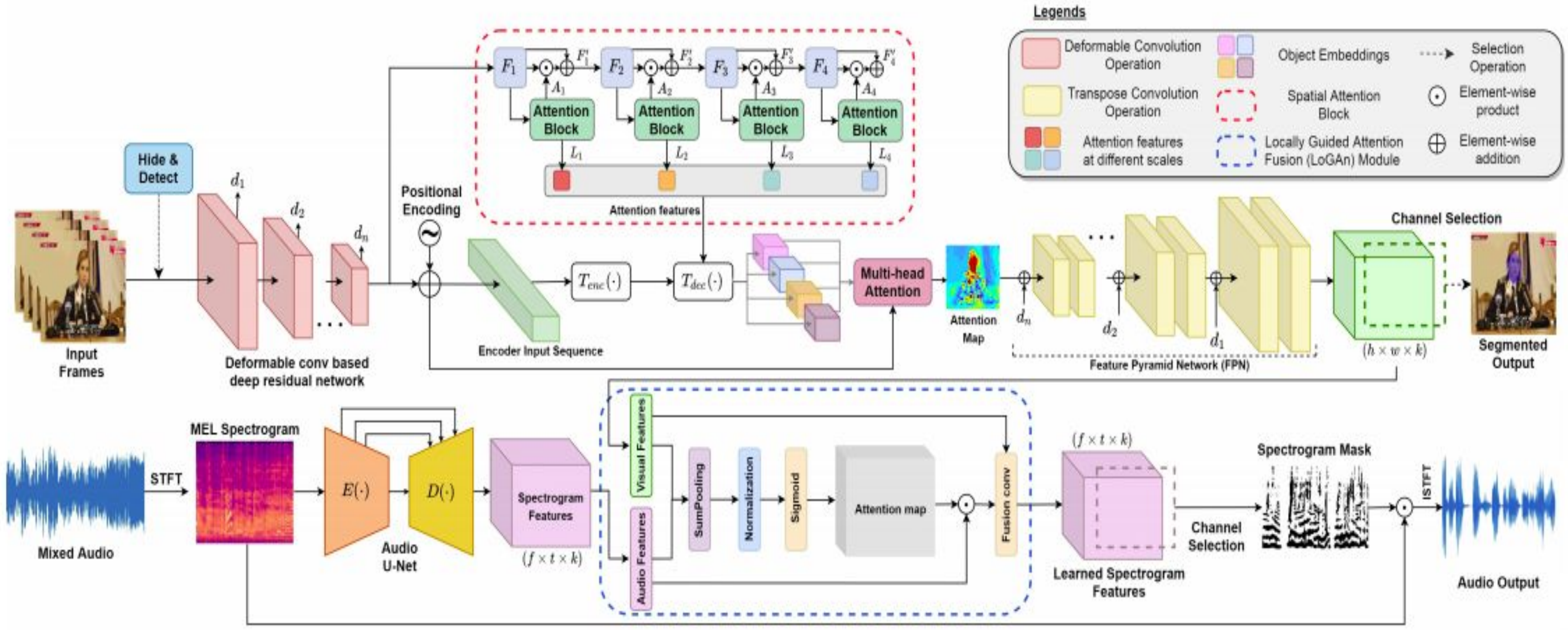
Our Proposed Work

- In this work we aim to solve the joint audio-visual segmentation problem in a self-supervised manner by leveraging the audio and visual modalities

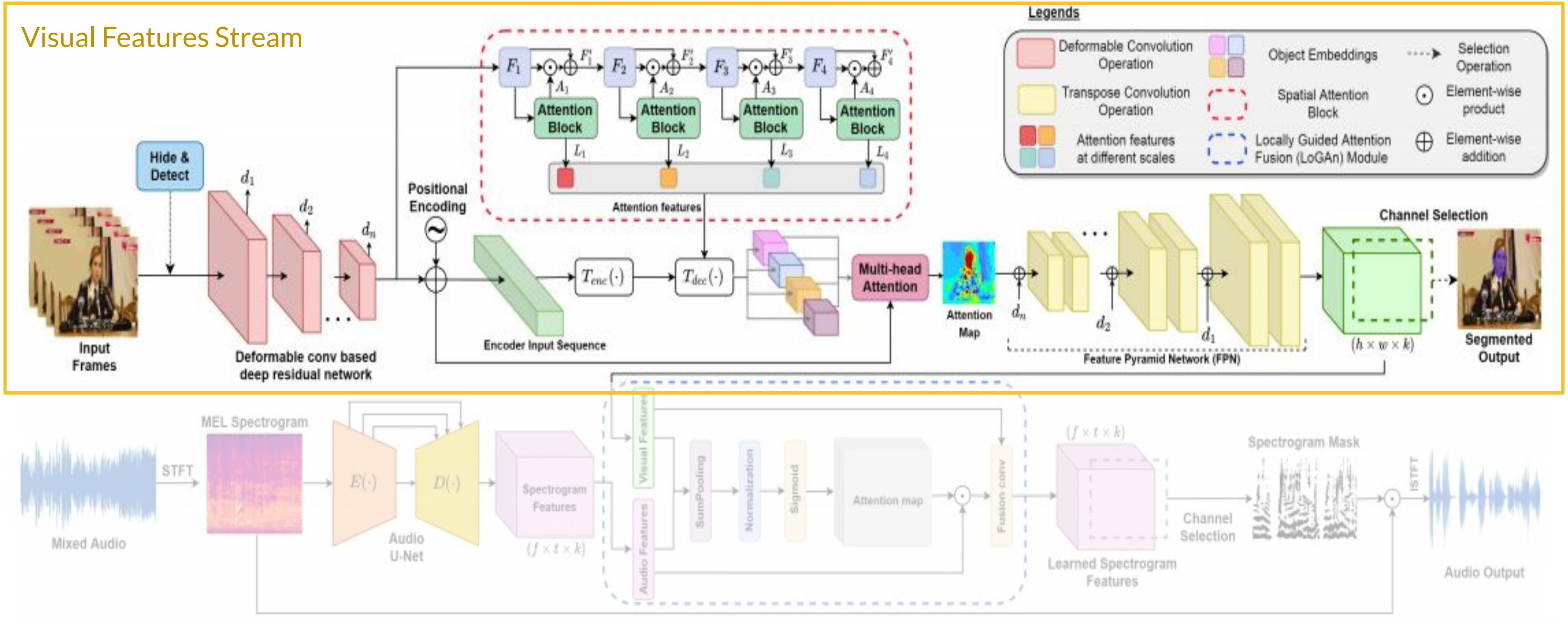
- Our network is able to blend cross-modal information more efficiently to extract the high level semantic information

- And more importantly, it works equally well even in cases of occluded sound source segmentation and also the segmentation of multiple but similar acoustic sources.

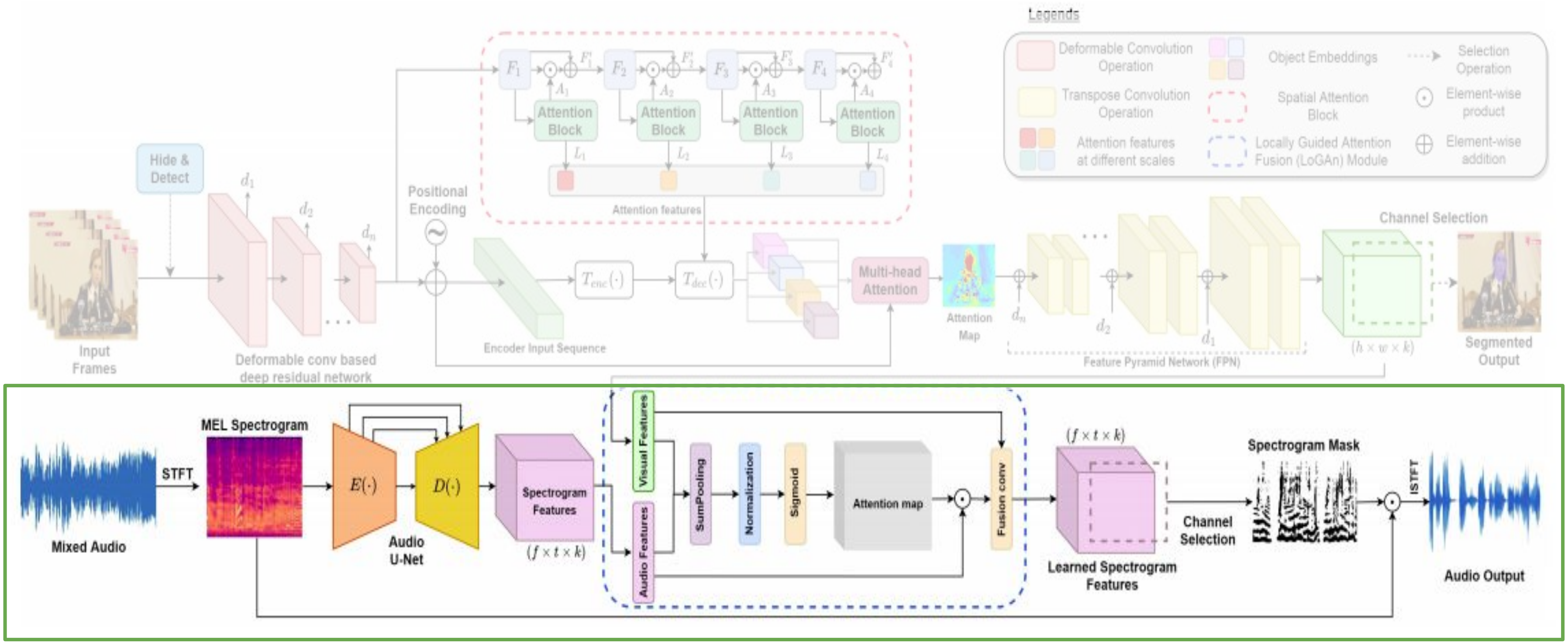
Architecture of the Proposed Network



Architecture of the Proposed Network



Architecture of the Proposed Network



Audio Features Stream

STREAM 1: Visual Features Stream

- The first input stream of the network aims to extract the Visual Features for the purpose of doing the segmentation with the help of audio signal
- We make use of a Deformable Convolution [2] based ResNet [1] backbone to extract a dense feature representation
- The visual segmentation path comprises a 'Spatial Attention Block' to enable a Transformer network based encoder-decoder to obtain an attention map of the sound-producing object(s)
- We use a Features Pyramid Network [3] to extract the multi-channel learnt attention based features to get the segmentation map.

[1]Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in CVPR, 2016, pp. 770–778.
[2] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei, "Deformable convolutional networks," in ICCV, 2017, pp. 764–773.
[3] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, et al., "Feature pyramid networks for object detection," in CVPR, 2017, pp. 2117–2125.

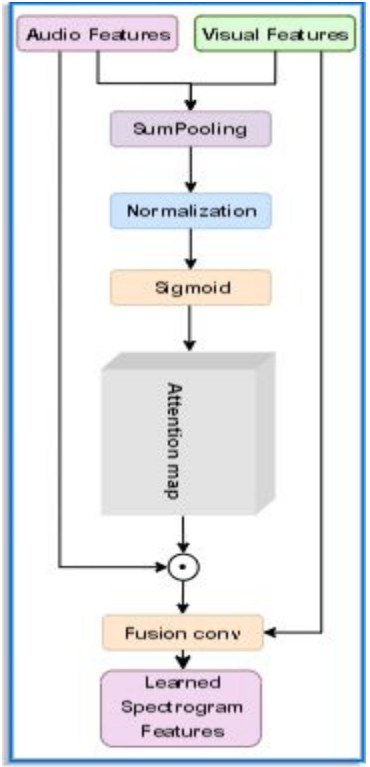
STREAM 2: Audio Features Stream

- To convert the audio signals into spectrograms, we use Short-Time Fourier Transform (STFT) with window size and hop length of 1022 and 256 respectively.

- Audio separation module performs feature extraction using an Audio U-Net that is later used along with the visual features for the sound source separation task.

Cross-modal learning through LoGAn

- The extracted audio and visual features need to be fused productively to facilitate cross-modal learning.
- LoGAn fusion module is used to allow high-level associations of audio and video features by capturing semantic information.
- The aggregated fusion feature map is obtained using a few convolutional layers over the visual features and the pixel-wise multiplication of the audio features and the attention map.



Cross-modal learning through LoGAn

- Attention map M_t is obtained by applying sum-pooling followed by *Power* and L_2 normalization
- Values in the attention map ranges between [0,1]; where 0 represents non-coherence between audio and visual cues and 1 represents high association
- The aggregate feature map F_{agg} can be formulated as:

$$F_{agg} = Conv([V_t, M_t \odot A'_t])$$

where, $[\cdot, \cdot]$ denotes concatenation operation

- This novel approach of cross-modal information blending turns out to be very efficient for the task
- We apply binary mask with per pixel sigmoid CE Loss

Partially occluded sound source capture

- ‘Hide-and-detect’ approach - mask the occluded source features before feeding it to the transformer encoder
- Curriculum learning strategy by initially masking the entire acoustic source
- Gradually masking smaller segments in order to train the network for the occluded source segmentation task

Audio guided segmentation

- We use audio information to segment multiple (but similar) sound sources present in the visual scene
- Audio intensity is faint for objects at greater depths
- We follow [1] to detect the presence of another instance of the same kind

(a) (b)

Fig. : Inference of AViS-Net: (a) without using audio information, (b) on using audio information.

[1] - Arthur N´adas, David Nahamoo, and Michael A Picheny, "Speech recognition using noise-adaptive prototypes," in ICASSP. IEEE, 1988, pp. 517–518.

Experimental results

Performance comparison with contemporary methods shows that individual components perform significantly well for audio-visual joint segmentation tasks. We consider Audio-Visual Event (AVE) [7] dataset for all the experiments.

Table 1: Performance comparison with respect to sound separation and semantic segmentation (IoU threshold 75%).

Method	**SDR**	**SIR**	**Visual Segmentation Accuracy (%)**
Audio feature only	5.28	9.43	59.68
Visual feature only	4.16	6.88	63.49
Zhao et al. [6]	1.03	6.37	45.90
PixelPlayer [5]	4.96	9.21	64.42
AViS-Net [ours]	**7.43**	**13.16**	**70.95**

[5] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba, "The sound of pixels," in ECCV, 2018, pp. 570–586.
[6] Andrew Rouditchenko, Josh McDermott, Antonio Torralba, et al., "Self-supervised audio-visual co-segmentation," in ICASSP. IEEE, 2019, pp. 2357–2361.
[7] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu, "Audio-visual event localization in unconstrained videos," in ECCV, 2018, pp. 247–263.

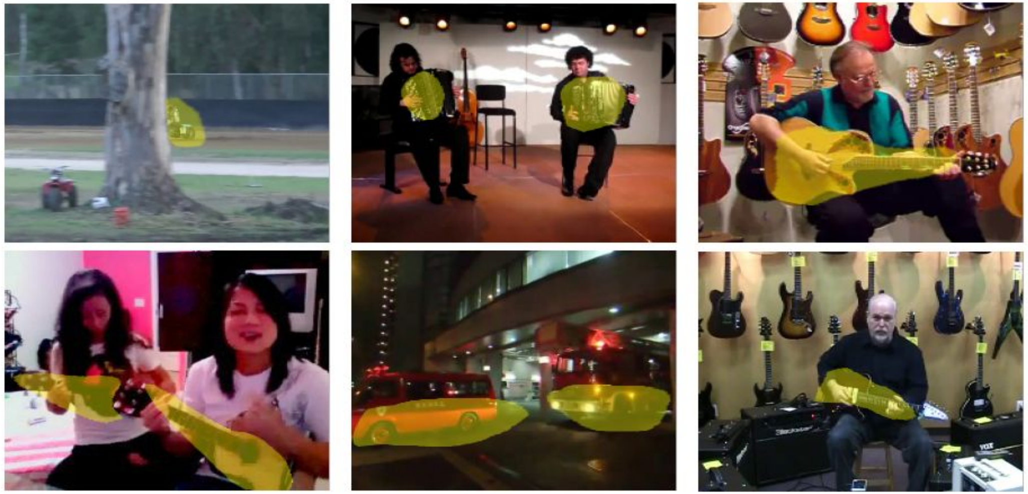
Experimental results

Following table shows the effectiveness of our novel fusion mechanism. The proposed feature fusion strategy has improved the overall performance by a considerable margin over existing methods of element-wise addition (EA) or element-wise multiplication (EM).

Table 2: Comparison of fusion strategies of audio and visual features (IoU threshold 75%).

Fusion Strategy	SDR	SIR	SAR	Visual Segmentation Accuracy (%)
EM	4.32	7.29	6.19	56.38
EA	5.11	8.24	7.22	59.96
Concatenation	5.99	9.38	9.03	64.13
LoGAn [ours]	**7.43**	**13.16**	**12.84**	**70.95**

Visual results



(a) (b) (c)

Fig. - Sound-source segmentation by AViS-Net:
(a) Partially occluded sound source,
(b) Multiple similar sound sources,
(c) Only one among multiple similar objects is producing sound.

Conclusion

- We leverage the concurrency between audio and visual modalities in an attempt to solve the joint audio-visual segmentation problem in a self-supervised manner.
- We propose a novel audio-visual fusion network, LoGAn, which captures high-level semantic information leading to superior performance.
- We are the first to address the partially occluded sound source segmentation task.
- In future, we plan to scale this task for more complex scenarios like ‘in-the-wild’ acoustic sources and more accurate segmentation & separation.

Thank You!

Sanjoy Chowdhury - schowdhury671@gmail.com
Subhrajyoti Dasgupta - subhrajyotidg@gmail.com
Sudip Das - d.sudip47@gmail.com
Ujjwal Bhattacharya - ujjwal@isical.ac.in