Listen To The Pixels

Paper #2351

IEEE ICIP 2021

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

[1]IIIT, Hyderabad, [1]ShareChat, Bangalore, [2] Indian Statistical Institute, Kolkata

PROBLEM

In this work, we leverage the concurrency between **Audio** and **Visual** modalities to solve the joint audio-visual segmentation problem in a **Self-supervised** manner.

Fig. 1: Illustration represents sequential, non-overlapping segments (shown by dotted lines) along with localization of the sound source and their corresponding segmented audio signals inferred by the proposed *AViS-Net* framework.

CHALLENGES

- Lack of annotated data
- Efficient blending of cross-modal information
- Partially occluded sound source segmentation etc.

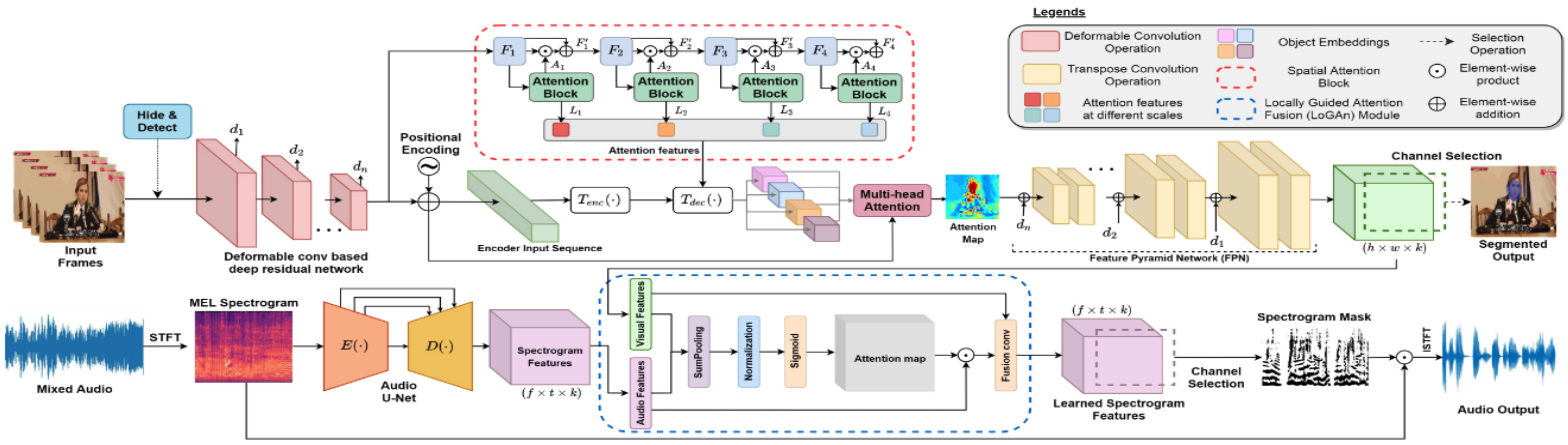


Fig. 2: An overview of AViS-Net architecture. The visual segmentation path comprises a transformer network based encoder-decoder that eventually leads to sound source segmentation. The audio separation module performs feature extraction using an Audio U-Net that is later used along with the visual features for the sound source separation task. Both the visual and audio features are fused using a LoGAn module.

1 Audio-Visual Segmentation Network (AViS-Net)

- The two-stream network takes both audio and visual data as inputs and exploits global and local event information efficiently to carryout cross-modal joint segmentation.
- Annotated data not required, follows **self-supervised** strategy

2 Cross-modal learning through Locally Guided Attention (LoGAn)

- Network needs to temporally adjust the audio and video feature maps at pixel level
- Applied binary masks with a per pixel sigmoid cross entropy loss, where the backpropagation facilitates cross-modal learning

3 Partially occluded sound source segmentation

- Hide-and-detect masks the occluded source features before feeding to the transformer encoder during training
- Curriculum learning strategy was deployed to address increasingly challenging examples

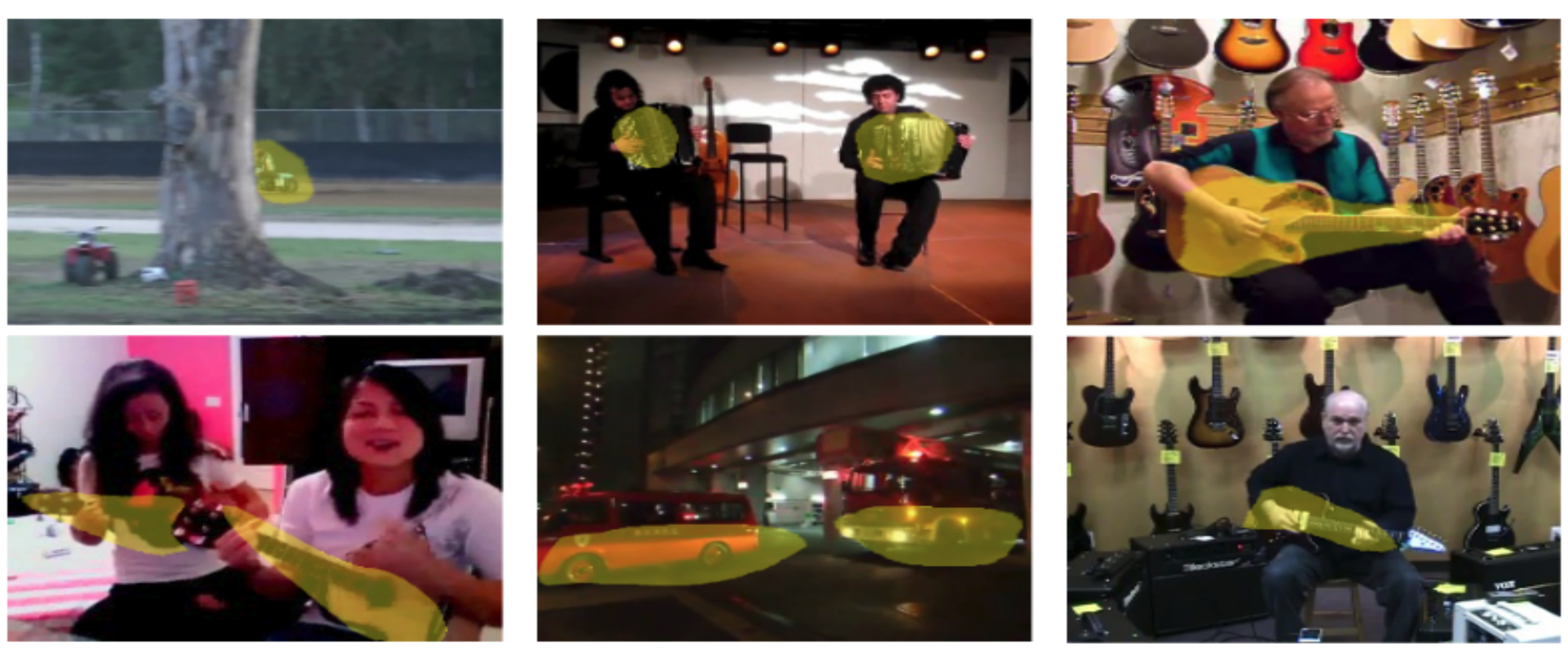
4 Audio guided segmentation

Exploit audio information to segment multiple (but similar) acoustic sources present in the visual scene

(a) (b)

Fig. 3: Inference of AViS-Net: (a) without using audio information, (b) on using audio information.

RESULTS



(a) (b) (c)

Fig. 4: 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.

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**

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**