Localize to Binauralize: Audio Spatialization from Visual Sound Source Localization
SUPPLEMENTARY MATERIAL

Kranthi Kumar Rachavarapu Aakanksha Vignesh Sundaresha Rajagopalan A. N.
Indian Institute of Technology Madras, India
{kranthi.rachavarapu, aakankshajha30, vigneshsundaresh}@gmail.com raju@ee.iitm.ac.in

S1. Qualitative comparison of Binaural Audio

The video results are available on our project page[1]. This file contains a compilation of videos for qualitative comparison. The following are the time-stamps of the results in this video:

- The localization output and the binaural audio generated using the proposed L2BNet trained with Weakly Semi-Supervised framework is available from $00:00-02:15$ sec.
- Comparison between the audio output of L2BNet generated using Weakly Supervised vs. Weakly Semi-Supervised framework is available from $02:15-04:03$ sec.
- A few samples used in User-study are available from $04:03-08:51$ sec.

S2. Additional Qualitative results for Sound Source Localization using Audio

Visual comparisons of Sound Source Localization using various input audio forms are available in Figure S1 and Figure S2.

S3. Qualitative results for Sound Source Localization in Weakly Supervised and Weakly Semi-Supervised Learning setup

Visual comparisons of Sound Source Localization task of the L2BNet train with *Weakly Supervised* setup is shown in Figure S3. SSL task output of L2BNet trained with Weakly Semi-Supervised setup with 10% supervision on FAIR-Play [1] dataset is shown in Figure S4 and on YTMusic [2] dataset is shown in Figure S5.

S4. Implementation Details

The proposed L2BNet consists of two subnetworks: Stereo-Generation Network and Audio Localization network. The layer-wise details of the SG network are reported in Table S1 and AL network are reported in Table S2.

References

[1] Ruohan Gao and Kristen Grauman. 2.5 d visual sound. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 324–333, 2019. 1

[2] Timothy Langlois Pedro Morgado, Nuno Vasconcelos and Oliver Wang. Self-supervised generation of spatial audio for 360deg video. In *Neural Information Processing Systems (NIPS)*, 2018. 1

[1] https://github.com/KranthiKumarR/Localize-to-Binauralize

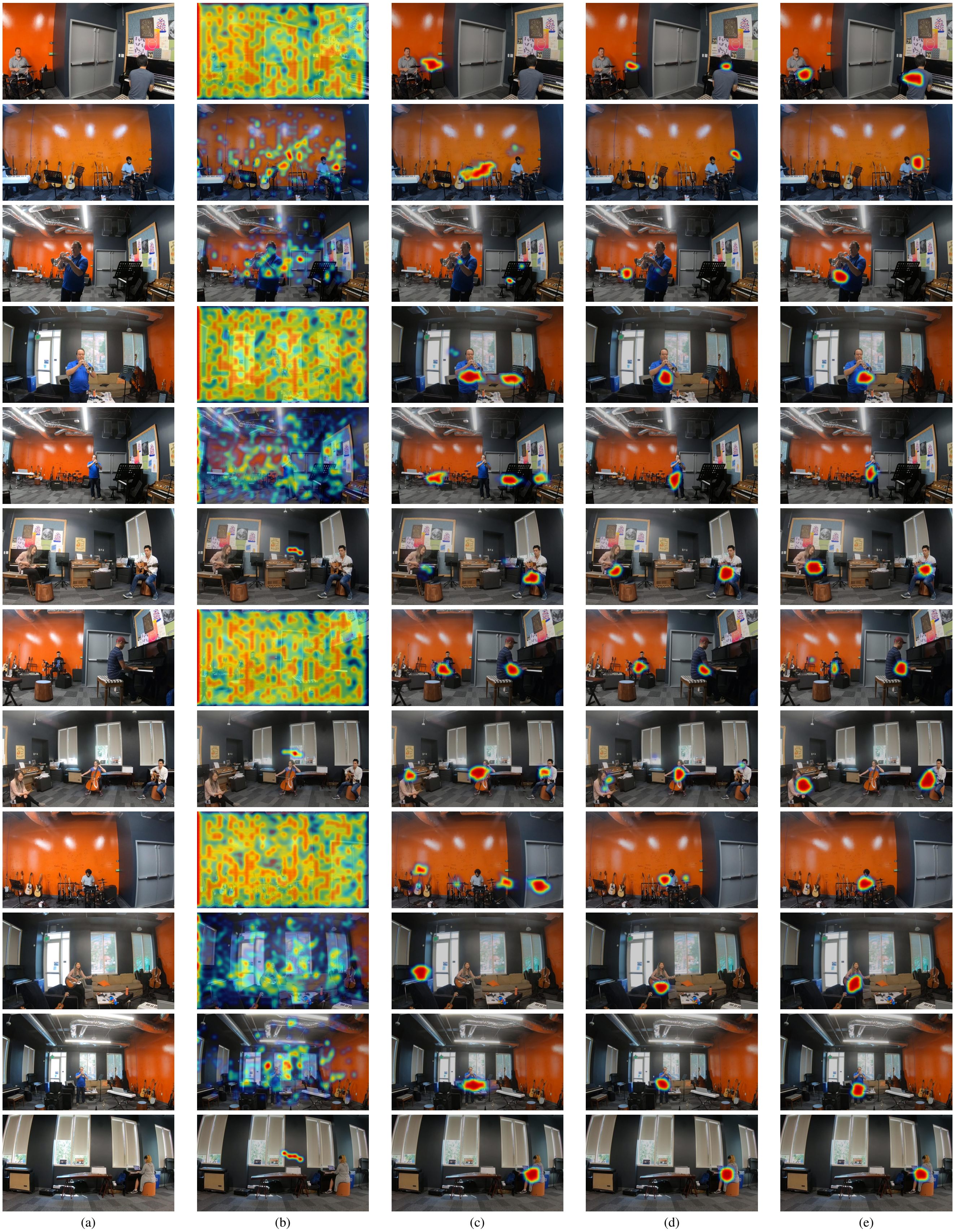
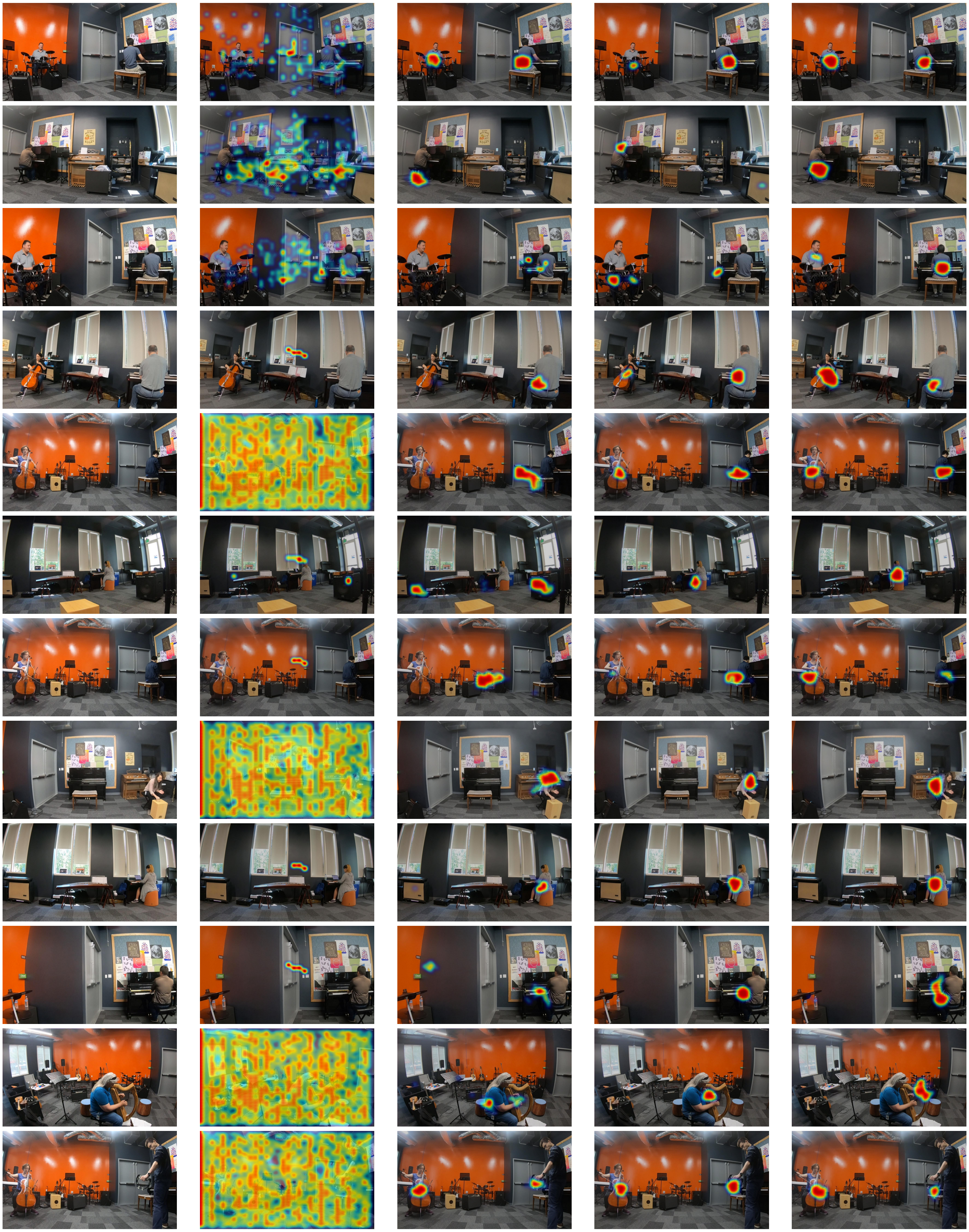
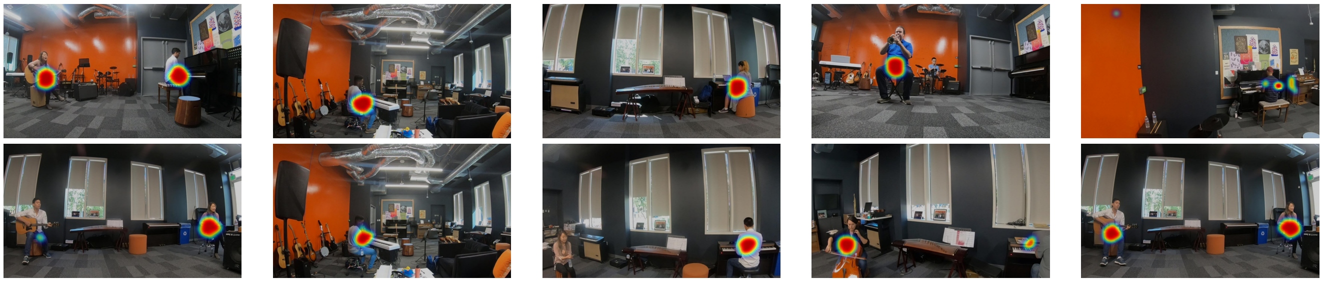


Figure S1. Visual comparisons of Audio-based Visual Sound Source Localization task using various input audio forms (a) *Visual frame* (b) from *monaural audio* (c) from *binaural mixed audio* (d) from *binaural audio* (e) *ILD&ITD* features.



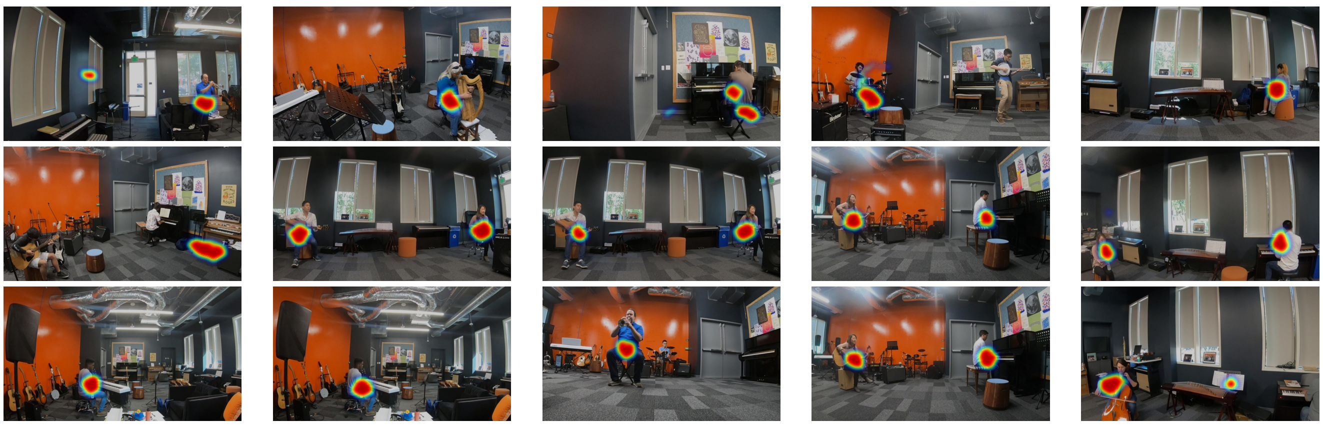
(a) (b) (c) (d) (e)

Figure S2. Visual comparisons of Audio-based Visual Sound Source Localization task using various input audio forms (a) *Visual frame* (b) from *monaural audio* (c) from *binaural mixed audio* (d) from *binaural audio* (e) *ILD&ITD* features.



(a) (b) (c) (d) (e)

Figure S3. Visual comparisons of sound source localization task of the L2BNet trained with *Weakly Supervised* learning setup.



(a) (b) (c) (d) (e)

Figure S4. Visual comparisons of sound source localization task of the L2BNet trained with *Weakly Semi-Supervised* learning setup.

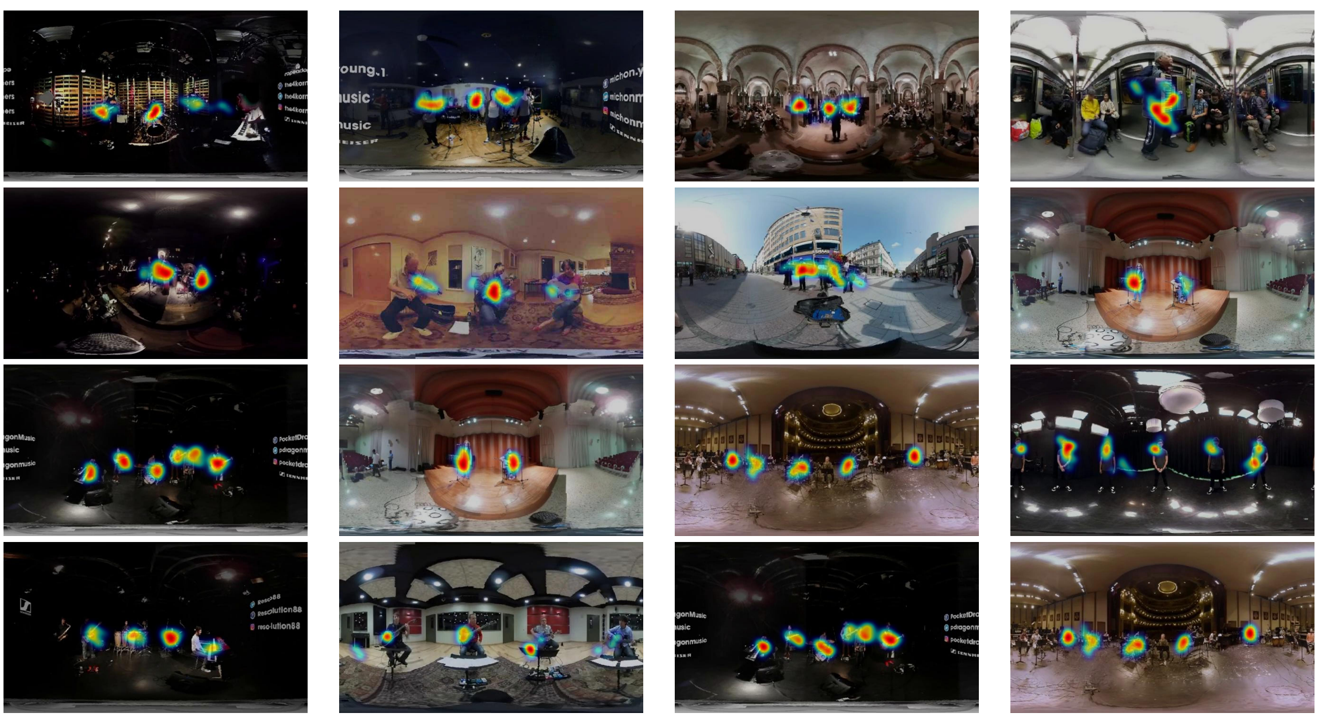


Figure S5. Visual comparisons of sound source localization task of the L2BNet trained with *Weakly Semi-Supervised* learning setup on YTMusic Dataset.

Subnetwork	Name	Type	K	S	Out
Audio Subnetwork	Encoder	Conv-2D	4	2	64
		BatchNorm-2D	-	-	64
		Conv-2D	4	2	128
		BatchNorm-2D	-	-	128
		Conv-2D	4	2	256
		BatchNorm-2D	-	-	256
		Conv-2D	4	2	512
		BatchNorm-2D	-	-	512
		Conv-2D	4	2	512
		BatchNorm-2D	-	-	512
	Decoder	ConvTranspose-2D	4	2	512
		BatchNorm-2D	-	-	512
		ConvTranspose-2D	4	2	256
		BatchNorm-2D	-	-	256
		ConvTranspose-2D	4	2	128
		BatchNorm-2D	-	-	128
		ConvTranspose-2D	4	2	64
		BatchNorm-2D	-	-	64
		ConvTranspose-2D	4	2	2
		BatchNorm-2D	-	-	2
Visual Subnework	Pretrained ResNet-18				
Attention	Query	Conv-2D	3	1	512
	Key	Conv-2D	3	1	512
	Value	Conv-2D	3	1	512

Table S1. Architecture summary of Stereo-Generation Network. K stands for kernel size, S for stride, n_f for number of input feature channels and *Out* for number of channels in convolutional layers. All the layers use *Leaky-ReLU* activation.

Name	Type	K	S	Out
Encoder	Conv-2D	4	2	64
	BatchNorm-2D	-	-	64
	Conv-2D	4	2	128
	BatchNorm-2D	-	-	128
	Conv-2D	4	2	256
	BatchNorm-2D	-	-	256
	Conv-2D	4	2	512
	BatchNorm-2D	-	-	512
	Conv-2D	4	2	1024
	BatchNorm-2D	-	-	1024
	AveragePool-2D			
Decoder	ConvTranspose-2D	4	2	512
	BatchNorm-2D	-	-	512
	ConvTranspose-2D	4	2	256
	BatchNorm-2D	-	-	256
	ConvTranspose-2D	4	2	128
	BatchNorm-2D	-	-	128
	ConvTranspose-2D	4	2	32
	BatchNorm-2D	-	-	32
	ConvTranspose-2D	4	2	4
	BatchNorm-2D	-	-	4
	ConvTranspose-2D	4	2	1
	BatchNorm-2D	-	-	1

Table S2. Architecture summary of Audio-Localization Network. K stands for kernel size, S for stride, n_f for number of input feature channels and *Out* for number of channels in convolutional layers. All the layers use *Leaky-ReLU* activation.