– Supplementary Material –
Less Can Be More: Sound Source Localization With a Classification Model

Here, we present additional details pertaining to the experiments that could not be included in the main text due to space constraints. All figures and references in this supplementary file are self-contained.

The contents included in this supplementary material are as follows: 1) The network architecture, 2) Video classification accuracy of our model on VGGSound dataset, 3) Difference between our pseudo bounding boxes and visual object detectors 4) Additional discussion and qualitative results for selected informative moments vs. mid-frames 5) Details for the modified version of [4] that is used in Section 5 of the main paper.

1. Architecture

In Table 1, we provide the architecture of the backbone networks. We use a two-stream network architecture that contains video and audio network, as in existing audio-visual learning works. The video network is a spatio-temporal ResNet mixed convolution network, similar to MCx [6], borrowed from official PyTorch implementation[1]. Audio network is a network that contains 2D convolution layers, similar to [1]. Batch Normalization and ReLU activation function are used after every convolution layer.

Layer	# filters	K	S	P	Output
input	1	-	-	-	$10 \times 100 \times 80$
conv1	64	(1,3,3)	(1,2,1)	(0,1,1)	$10 \times 50 \times 80$
conv2	64	(1,3,3)	(1,1,2)	(0,1,1)	$10 \times 50 \times 40$
maxpool2	-	(1,1,3)	(1,1,2)	(0,0,0)	$10 \times 50 \times 19$
conv3	192	(1,3,3)	(1,1,1)	(0,1,1)	$10 \times 50 \times 19$
maxpool3	-	(1,3,3)	(1,2,2)	(0,0,0)	$10 \times 24 \times 9$
conv4	256	(1,3,3)	(1,1,1)	(0,1,1)	$10 \times 24 \times 9$
conv5	256	(1,3,3)	(1,1,1)	(0,1,1)	$10 \times 24 \times 9$
conv6	256	(1,3,3)	(1,1,1)	(0,1,1)	$10 \times 24 \times 9$
maxpool6	-	(1,3,2)	(1,2,2)	(0,0,0)	$10 \times 11 \times 4$
conv7	512	(1,4,4)	(1,1,1)	(0,1,0)	$10 \times 10 \times 1$
fc8	512	(1,1,1)	(1,1,1)	(0,0,0)	100×1
fc9	512	(1,1,1)	(1,1,1)	(0,0,0)	100×1

(a) Audio Network

Layer	# filters	K	S	P	Output
input	3	-	-	-	$100 \times 112 \times 112$
conv1	64	(3,7,7)	(1,2,2)	(1,3,3)	$100 \times 56 \times 56$
conv2	64	(3,3,3)	(1,1,1)	(1,1,1)	$100 \times 56 \times 56$
conv3	64	(3,3,3)	(1,1,1)	(1,1,1)	$100 \times 56 \times 56$
conv4	64	(3,3,3)	(1,1,1)	(1,1,1)	$100 \times 56 \times 56$
conv5	64	(3,3,3)	(1,1,1)	(1,1,1)	$100 \times 56 \times 56$
conv6	128	(1,3,3)	(1,2,2)	(0,1,1)	$100 \times 28 \times 28$
conv7	128	(1,3,3)	(1,1,1)	(0,1,1)	$100 \times 28 \times 28$
res-conv8	128	(1,1,1)	(1,2,2)	(0,0,0)	$100 \times 28 \times 28$
conv9	128	(1,3,3)	(1,1,1)	(0,1,1)	$100 \times 28 \times 28$
conv10	128	(1,3,3)	(1,1,1)	(0,1,1)	$100 \times 28 \times 28$
conv11	256	(1,3,3)	(1,2,2)	(0,1,1)	$100 \times 14 \times 14$
conv12	256	(1,3,3)	(1,1,1)	(0,1,1)	$100 \times 14 \times 14$
res-conv13	256	(1,1,1)	(1,2,2)	(0,0,0)	$100 \times 14 \times 14$
conv14	256	(1,3,3)	(1,1,1)	(0,1,1)	$100 \times 14 \times 14$
conv15	256	(1,3,3)	(1,1,1)	(0,1,1)	$100 \times 14 \times 14$
conv16	512	(1,3,3)	(1,2,2)	(0,1,1)	$100 \times 7 \times 7$
conv17	512	(1,3,3)	(1,1,1)	(0,1,1)	$100 \times 7 \times 7$
res-conv18	512	(1,1,1)	(1,2,2)	(0,0,0)	$100 \times 7 \times 7$
conv19	512	(1,3,3)	(1,1,1)	(0,1,1)	$100 \times 7 \times 7$
conv20	512	(1,3,3)	(1,1,1)	(0,1,1)	$100 \times 7 \times 7$
avgpool	-	(1,7,7)	-	(0,0,0)	$100 \times 1 \times 1$

(b) Video Network

Table 1: **Architecture of the backbone networks.** K, S, P, res, $maxpool$ and $avgpool$ denote kernel size, stride, padding, residual, max-pooling and average-pooling layers, respectively.

2. Video Classification Results

The proposed model is trained with the objective of video classification. Here, we show experimental results for predicting video-level labels with our proposed model and the other model settings, presented in the ablative study. Table 3 shows the video classification accuracies. As results indicate, the proposed model gives the highest classification performance which is aligned with the findings of [6] and the trend of sound localization results in "Ablative Study" of the main paper.

3. Automatic Bounding Box Generation for Sounding Objects

In Section 5.2 of the main paper, we show that our proposed model has an ability to automatically generate accurate pseudo bounding boxes for sounding objects. In this section, we give a more detailed discussion.

First, even though our qualitative results show accurate boxes for sounding objects in the scene, one can raise a question about the difference between the usage of off-the-shelf visual detectors and our results. To show the effec-

[1] https://pytorch.org/vision/0.8/models.html#torchvision.models.video.mc3_18

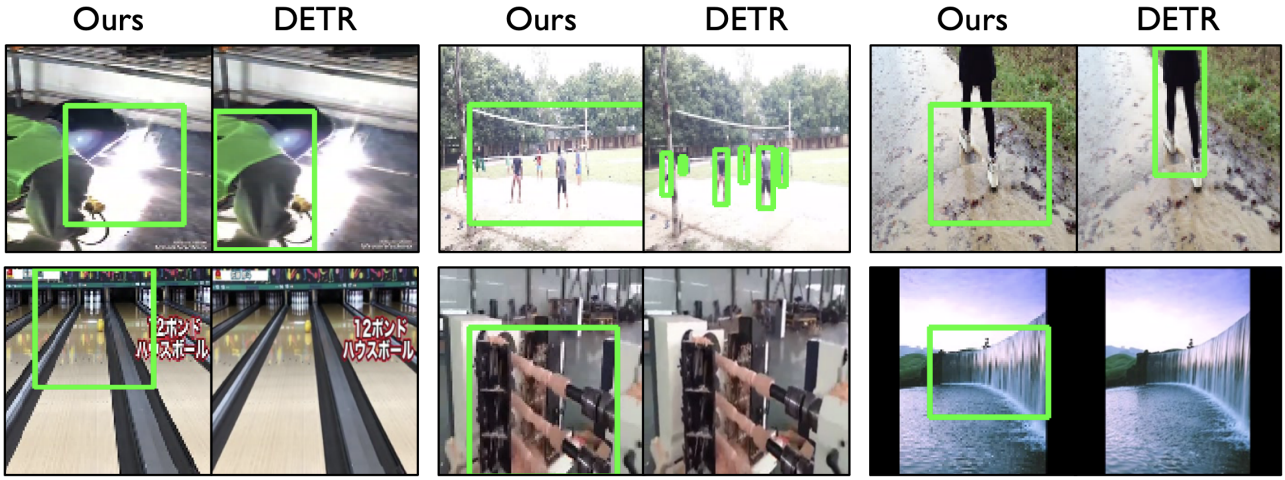

Figure 1: **Qualitative Results of Automatic Bounding Box Generation and comparison with DETR [3].** Our method accurately generates bounding boxes for sounding objects to be used in sub-tasks whereas off-the-shelf visual detector DETR [3] can not propose proper boxes.

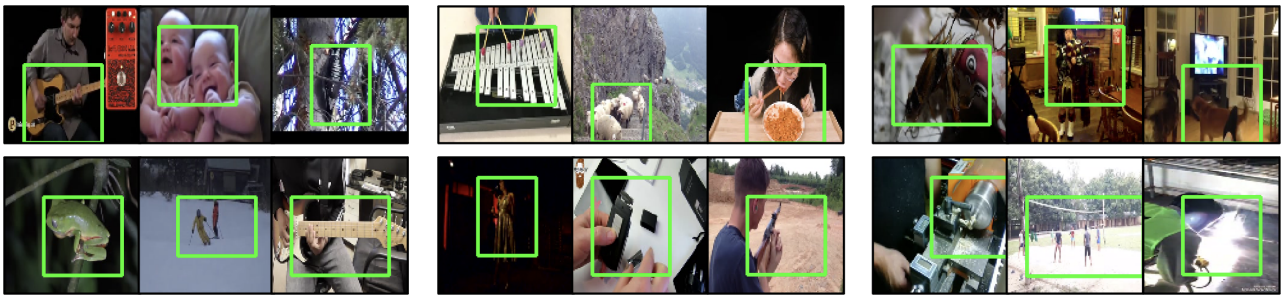

Figure 2: **Qualitative Results of Automatic Bounding Box Generation.** Our method accurately generates bounding boxes for sounding objects to use in sub-tasks, such as faster human annotation and semi-supervised sound localization.

Method	AV-head	A-head	V-head
Single Multi-Modal	57.1	–	–
Shared FC	57.6	45.6	39.1
Individual FCs	**58.1**	**45.9**	**39.2**

Table 3: **Video classification performance on VGGSound dataset**. We investigate the effects of architectural choices in the proposed method. AV-head, A-head, and V-head indicate audio-visual head, audio-head, and visual-head respectively.

tiveness of our results, we make a comparison with a recent visual detector [3]. Figure 1 shows this comparison. As it can be seen in "welding" example (first column of the first row in Figure 1), visual detector (DETR) has no ability to indicate that sounding location in the scene is the welding light. Also, "volleyball" example shows that our network outputs a box around the volleyball field including people. However, the visual detector only focuses on humans individually. Additionally, we can see that visual detector can not propose any bounding box for some examples. The second row of the Figure 1 shows that DETR misses some locations or objects in the scene while our proposed method can give accurate pseudo boxes for possible sound locations.

Lastly, our method generates accurate bounding boxes not only for the samples that contain a single object with a relatively simple background but also in complex scenes. As it can be seen in the last column of the first row in Figure 2, our network proposes a bounding box around the dog even though there are furniture and TV in the scene. More qualitative results for the ability of our method are presented in Figure 2.

4. Additional Results of the Informative Moment Selection

As shown in the main paper, the sound localization task performance gets higher with the properly selected moments. To justify our idea, we visualize more samples that have a big difference between the mid-frame and the selected moment. Largely we can categorize the samples as audio-wise and visual-wise wrong samples. We visualize

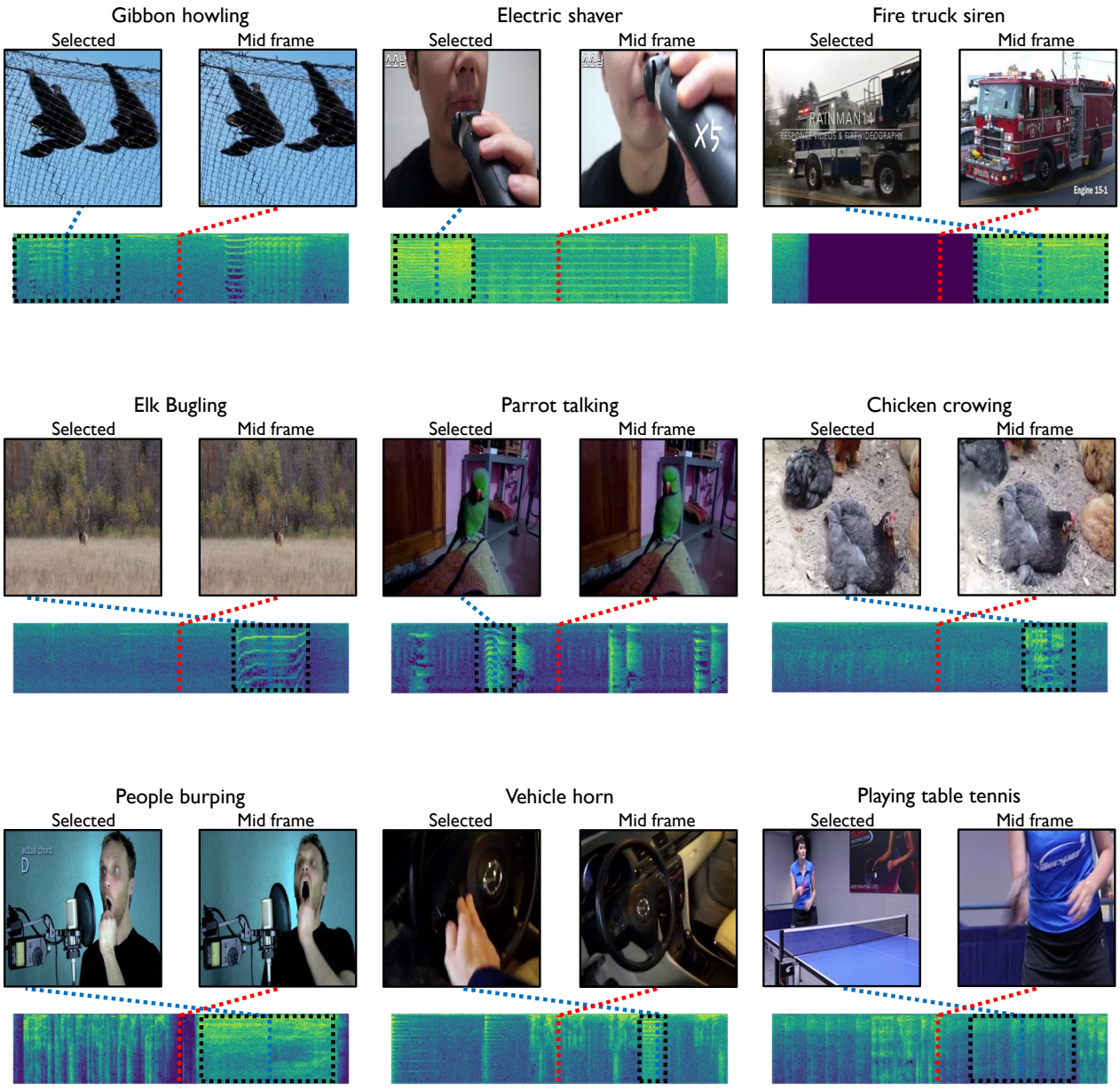

Figure 3: **Audio-wise comparison of mid-frame and our informative moment selection strategies.** The red dashed lines are the mid-frame moments, and the blue depicts our selected moments. The black dashed boxes contain the proper information in the audio modality. Mid-frame selection has drawbacks as picking a non-informative part of the audio signal whereas our proposed strategy gives more useful inputs for sound localization training. For example; "fire truck siren" and "elk bugling" examples show that our proposed strategy selects the moment that has an informative audio signal.

the spectrogram for audio-wise wrong samples along with the corresponding visual frames (Figure 3). The sound of the beast howling or crowing can be considered as an instant/sudden event. Similarly, fireworks or explosions show a visually meaningful moment within a short duration. (Figure 4). Some sub-sequences of videos lose audio-visual correspondence due to the manipulation of the videos such as clipping, concatenation, narration or mute. Some parts of videos contain visually noisy contents due to camera works such as defocusing or fading in and out. Because of the reasons mentioned above, selecting a mid-frame as a training sample for audio-visual tasks lower the chance to correlate the audio and the corresponding visual signal. In order to prevent the model learn from wrong correspondences, moment selection is important.

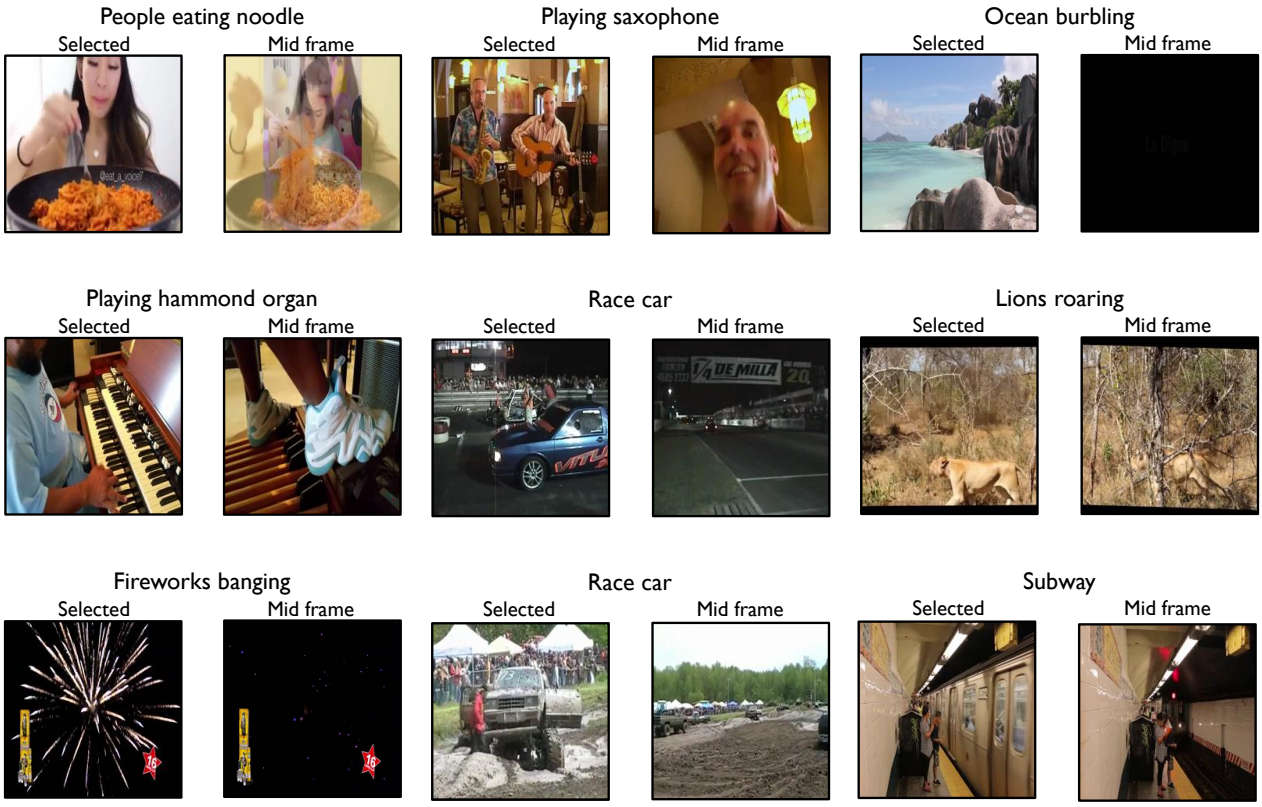

Figure 4: **Vision-wise comparison of mid-frame and our informative moment selection strategies.** Because of scene changes or object occlusion or sudden/instant events, the mid-frame strategy can pick non-informative visual frames. However, our proposed method can select audio-visually correlated informative visual frames.

5. Details of Modified Sound Localization Model

As it is mentioned in the main paper, we use the modified version of the publicly available sound localization network [4]. In [4], VGG-16 [5] is used for the vision embedding and SoundNet [2] is used for the audio embedding. As the backbone networks used in [4] are outdated, we replace them with ResNet-18 models. Following that backbone networks in [4] use pre-trained models for both vision and audio embedding, we also use ImageNet pre-trained ResNet-18 for the vision backbone. Since there is no available pre-trained audio network (ResNet-18) trained on VGGSound, we train our audio backbone network from scratch by following the training scheme of SoundNet, which learns the audio representation by knowledge distillation from the pre-trained vision network. The rest of the architecture and the training settings of [4] are identically followed.

References

[1] Triantafyllos Afouras, Andrew Owens, Joon Son Chung, and Andrew Zisserman. Self-supervised learning of audio-visual objects from video. In *European Conference on Computer Vision*, 2020.

[2] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. *Advances in Neural Information Processing Systems*, 2016.

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, 2020.

[4] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[5] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[6] Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.