# Exploiting Visual Context Semantics for Sound Source Localization Supplementary Material

Xinchi Zhou[1*] Dongzhan Zhou[1*] Di Hu[2] Hang Zhou[3] Wanli Ouyang[1]
[1]The University of Sydney [2]Gaoling School of Artificial Intelligence, Renmin University of China [3]Baidu Inc.
{xinchi.zhou1,d.zhou,wanli.ouyang}@sydney.edu.au,
dihu@ruc.edu.cn, zhouhang09@baidu.com

## A. Details on Experimental Settings

### A.1. Data Processing

For each video clip in the dataset, we take the middle frame and 3s audio segments around it to create the audio-visual pairs. The raw audio signals are re-sampled at 24k Hz and clipped to ensure their values are between -1 and 1. The signals are then converted to $257 \times 300$ magnitude spectrograms via STFT as the inputs to the audio extractor. The input frames are firstly resized to $256 \times 256$ and randomly cropped to $224 \times 224$ during training.

### A.2. Model Configurations

For both the audio and visual streams, we use a lightweight ResNet-18 [3] as the feature extractors. The final pooling and fully-connected layers are removed. Both the extractors are trained from **scratch** with random initialization instead of inheriting the ImageNet [6] pretrained weights as in [1]. In other words, the training process does not need any class labels. For the visual encoder, the stride of the last block is adjusted to 1 and the shape of the visual feature is $512 \times (H_0/16) \times (W_0/16)$, where $H_0$ and $W_0$ denote the height and width of input frames. For the audio side, we conduct a max-pooling operation to convert the audio output to a 512-$D$ vector. If not specified, we set the region number $N$ to 16 in the visual reasoning module, which consists of 8 ($N_P$) positive maps and 8 ($N_Q$) negative maps.

### A.3. Categories of VGGSound-MI Subset

The categories in the VGGSound-MI subset is shown in Table. A1.

### A.4. AudioSet Subset

We conduct training and testing on the subset of the AudioSet dataset, which includes 15 musical instruments.

| | | |
|---|---|---|
| playing accordion | playing acoustic guitar | playing banjo |
| playing bass drum | playing bass guitar | playing bassoon |
| playing bongo | playing cello | playing clarinet |
| playing congas | playing cornet | playing cymbal |
| playing djembe | playing double bass | playing drum kit |
| playing electric guitar | playing electronic organ | playing flute |
| playing french horn | playing gong | playing hammond organ |
| playing harmonica | playing harp | playing harpsichord |
| playing mandolin | playing oboe | playing piano |
| playing saxophone | playing snare drum | playing tabla |
| playing tambourine | playing timbales | playing timpani |
| playing trombone | playing trumpet | playing tympani |
| playing ukulele | playing violin, fiddle | tapping guitar |

Table A1. Musical instrument categories in VGGSound-MI subset.

The categories of instruments are: accordion, banjo, cello, drum, flute, french horn, guitar, harmonica, harp, oboe, piano, saxophone, trombone, trumpet, violin. Following [4], we use the 'unbalanced' split for training and the 'balanced' split for testing. The training set contains around 50k clips while the testing set contains 454 clips. Since there are no officially provided localization annotations, we follow the method suggested by [2] to generate the bounding boxes for evaluation. Specifically, we train an object detector using 30k images of the 15 categories from the OpenImage dataset [5]. The pre-trained detector is then used to extract bounding boxes of instruments in the AudioSet dataset. For the sake of precision, we only keep bounding boxes with confidence larger than 90%.

## B. Additional Visualization Results

In Figure. A1, we provide more qualitative results on the VGGSound test set to further verify the robustness of our approach. The results demonstrate that our method can achieve more precise localization results while LVS [1] exhibits more diffused predictions, leading to larger uncertain regions. We also find that the existence of 'people' may also interfere with the localization prediction, even though they are not the sources of sound. As shown in the first three rows of Figure. A1, LVS tends to incorporate hu-

*Equal contribution.

mans into the predictions while our method focuses on the sounding objects. The possible reason is that people appear most frequently in the training samples and LVS with only audio-visual correlation supervision fails to resist such disturbance. Conversely, our method enjoys supervisions from both the visual context semantics and the audio-visual correlations, which may alleviate the negative impacts.

## References


[1] Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Localizing visual sounds the hard way. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16867–16876, 2021.

[2] Ruohan Gao and Kristen Grauman. Co-separating sounds of visual objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3879–3888, 2019.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[4] Di Hu, Rui Qian, Minyue Jiang, Xiao Tan, Shilei Wen, Errui Ding, Weiyao Lin, and Dejing Dou. Discriminative sounding objects localization via self-supervised audiovisual matching. *Advances in Neural Information Processing Systems*, 33:10077–10087, 2020.

[5] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Andreas Veit, et al. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github.com/openimages*, 2(3):18, 2017.

[6] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

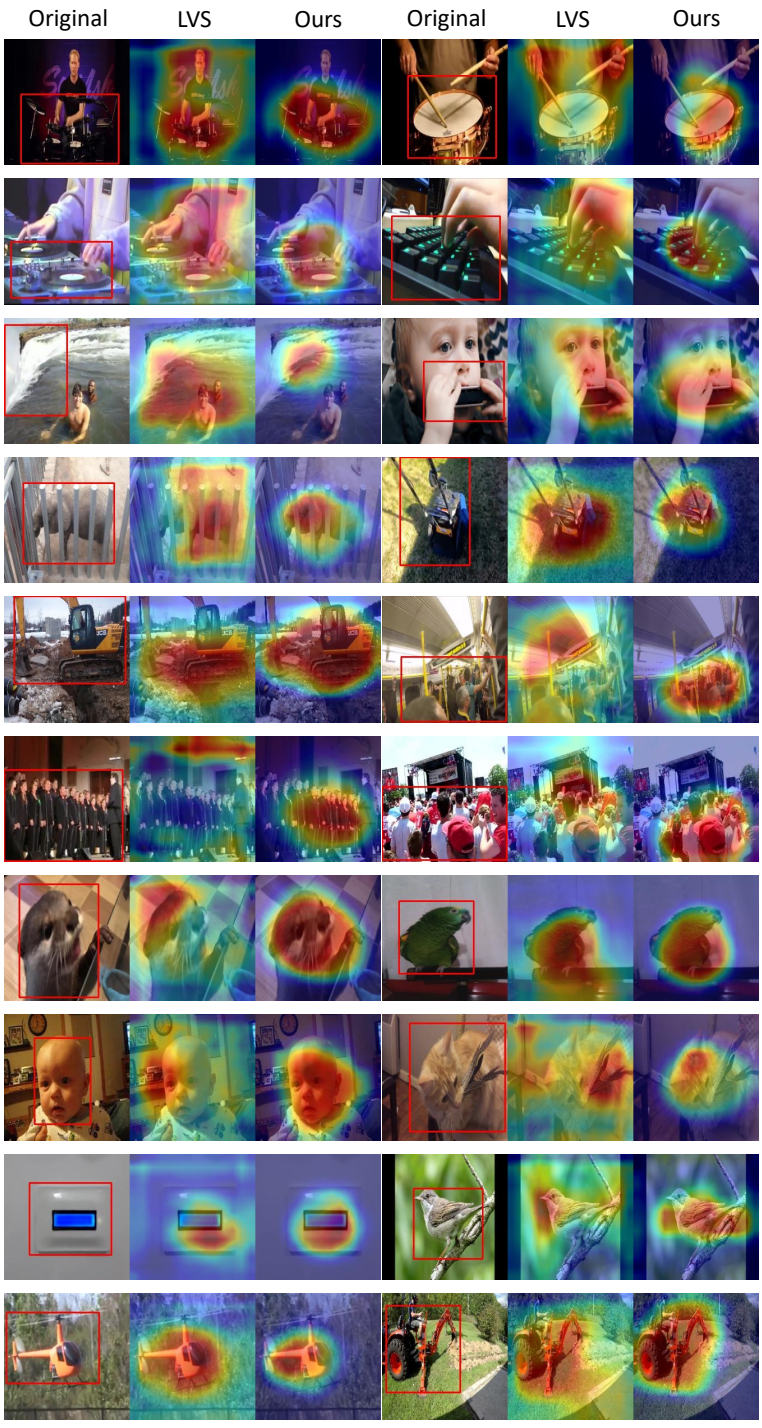


Figure A1. Visualization results on VGGSound test set. We visualize the localization maps between our approach and LVS [1]. The sounding objects are marked by the red bounding boxes in the original images.