Learning Audio-Visual Source Localization via False Negative Aware Contrastive Learning Supplementary Material

Weixuan Sun[1,5] *, Jiayi Zhang[2] *, Jianyuan Wang[3], Zheyuan Liu[1], Yiran Zhong[4], Tianpeng Feng[5], Yandong Guo[5], Yanhao Zhang[5], Nick Barnes[1]
[1]Australian National University, [2]Beihang University, [3]The University of Oxford, [4]Shanghai AI Lab, [5]OPPO Research Institute.

1. Implementation Details of Pilot Experiments

As mentioned in the main paper, we conducted a pilot experiment to investigate the false negative issue. In this section, we will introduce the implementation details of the pilot experiment. Considering different sounds of *playing drums* (as depicted in Fig. 1 of the main paper) are semantically matched and should share equal roles in locating the drums in visual scenes, we reasonably assume that samples from the same categories are mutually False Negatives.

Firstly, we examine the distribution of false negatives when training on the VGGSound [2] dataset. Specifically, we count the number of samples from the same category in a mini-batch, denoted as N. Then the ratio of false negatives in a mini-batch is calculated by N/B with batch size as B. When adopting random sampling with a batch size of 128, the mean ratio is statistically 39.27%. The ratio will undoubtedly increase when employing a bigger batch size, *e.g.*, 61.17% with a batch size of 256.

Secondly, we examine whether more false negatives will affect audio-visual localization performance. Concretely, we adopt ResNet-18 as the backbone to encode audio and visual features and a standard NCE [5] loss is used to update network parameters. When loading data of a training batch, we randomly select a category and then sample from the selected category with a probability of p. Since we randomly sample with the remaining probability of $1-p$, the proportion of false negatives in a mini-batch should be positively related to p and is always higher than p. Without loss of generality, we adopt a batch size of 128 and progressively increase the value of p from 0 to 0.5. Correspondingly, the actual ratio of false negatives increases from 39.27% to 60.79%. As shown in Table S1, a significant performance drop is observed, indicating that false negatives substantially harm the model quality.

*Indicates equal contribution

Table S1. Pilot experiment results on VGGSound-144k. FN Ratio represents the statistical ratio of false negatives during training.

p	FN Ratio(%)	Flickr CIoU(%)	VGG-SS CIoU(%)
0.0	39.27	79.93	36.11
0.1	43.21	78.72	34.95
0.2	46.86	78.32	34.91
0.3	51.37	76.71	33.98
0.4	55.83	75.91	33.94
0.5	60.79	67.48	25.42

Table S2. Ablation of warm up epochs.

warm-up	Flickr CIoU(%)	VGG-SS CIoU(%)
0	84.33	35.07
3	85.74	37.29
5	84.33	35.93
10	83.53	37.61

2. Ablation Analysis

In this section, we conduct thorough ablation experiments on various design choices in FNAC including warm-up epochs, dropout rate, and regularization distance loss, etc. All experiments are trained on VGG-Sound 10k and tested on Flickr and VGG-SS test sets.

2.1. Analysis of Warm-up Epochs

As introduced in the main paper, we warm up the network with only NCE loss for 3 epochs, then integrate our regularization for the remaining epochs during the training. We show that warm-up brings slightly better performances but is not necessary for FNAC. In Table S2, we ablate different warm-up epochs. As shown, 3 epochs of warm-up bring 1.41% CIoU improvement over the model without warm-up. Further, it is observed that a longer warm-up period does not bring additional performance gain. It indicates that FNAC does not rely on warm-up to obtain superior performances.

Table S3. Ablation of visual dropout rate.

Visual Dropout	Flickr CIoU(%)	VGG-SS CIoU(%)
0.1	77.10	35.76
0.3	81.12	34.75
0.5	82.32	36.43
0.7	84.73	36.49
0.9	85.74	37.29

Table S4. Ablation of different distance metrics for regularization loss. All models are trained on the VGG-Sound 10k training set.

Loss	Flickr CIoU(%)	VGG-SS CIoU(%)
L2	80.72	34.02
Smooth L1	82.32	34.47
L1	85.74	37.29

Table S5. Ablation of audio augmentation

Audio Aug	Flickr CIoU(%)	VGG-SS CIoU(%)
w augmentation	82.73	38.64
w/o augmentation	85.74	37.29

Table S6. Ablation of sounding region visual feature threshold. soft: no threshold is applied, the localization map is directly applied as a soft mask to extract localized visual features.

Thresh	Flickr CIoU(%)	VGG-SS CIoU(%)
0.2	81.92	37.74
0.3	83.53	38.32
0.4	82.73	37.70
0.5	83.53	37.29
0.6	85.74	37.29
0.7	84.73	34.49
Soft	83.13	36.00

2.2. Analysis of Visual Dropout Rate

An over-fitting issue is observed during the training which normally requires an early stopping strategy [1, 3, 4, 7]. To relieve the over-fitting issue, following the convention, we adopt dropout on the visual encoder and audio encoder. Empirically, we found that audio dropout does not affect the performances. However, visual dropout can effectively address the over-fitting and improve the results. We ablate different visual dropout rates in Table S3. As shown, we observe that the visual dropout is essential for performance and a larger dropout rate (0.9) achieves the best result. We hypothesize that it is because the audio features are relatively simple but the visual features are redundant. Thus, an audio clip tends to over-fit to the most discriminative visual region of its paired image and ignore all other visual entities, which hinders the ability of semantic-aware audio-visual localization. The visual dropout may encourage the audio features to be compared with broader visual features so as to increase localization robustness.

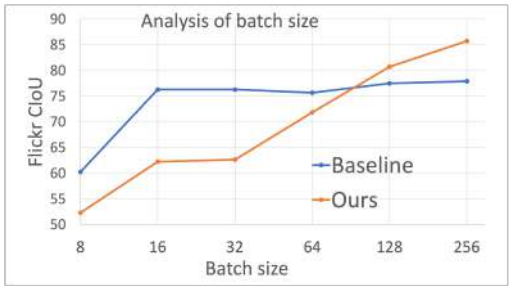

Figure S1. Analysis of batch size on Flick. Our method boosts performances when batch size increases.

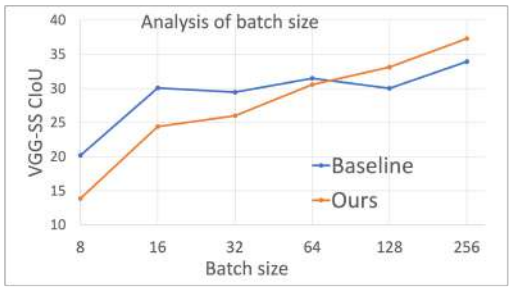

Figure S2. Analysis of batch size on VGG-SS. Our method boosts performances when batch size increases.

2.3. Analysis of Regularization Loss

In our implementation, we measure the distance between the adjacency matrices to enforce regularization. In Table S4, we ablate different distance metrics including l1 distance, l2 distance and smoothed l1 distance. As shown, using L1 distance achieves the best performance.

2.4. Analysis of Batch Size

In this section, we investigate the effects of batch sizes. As discussed in the main paper, more false negatives will be encountered in a large batch, whereas a larger batch size is preferred in contrastive learning for better representation [5]. We show results with different batch sizes in Fig. S1 and S2. As shown, when the batch size is 64 or less, the false negative ratio is low and nearly insignificant, so our method provides no benefit or even hinders performance. On the other hand, we observe a slight improvement in the baseline as batch size increases. Nevertheless, false negatives become prominent in large batch sizes, which is the focus of our paper. Our proposed method improves performance by effectively addressing false negatives, outperforming the baseline by a clear margin.

2.5. Analysis of Audio Augmentation

In this section, we ablate the audio augmentation. In our implementation, we adopt audio augmentations including Frequency mask and Time mask proposed in [6]. In Table S5, we report results with and without audio augmentation. As shown, we can see that audio augmentation brings improvement on Flickr test set but does not facilitate on the more challenging VGG-SS test set.

2.6. Analysis of Sounding Visual Feature Threshold

In TNE, we generate audio-visual localization maps and use the localization results to extract the localized visual features, i.e., sounding region visual representations. We enforce the extracted visual representations to be different between the true negative samples, which encourages the localization of authentic sound sources. In practice, we choose a threshold to separate sounding regions and quiet regions, lower threshold indicates stronger TNE since larger visual regions are considered in the training. We ablate this threshold in Table S6. We observe that the Flickr test set prefers a higher threshold (0.6) while VGG-SS prefers a lower threshold (0.3). We believe the reason for this is that the VGG-SS test set is more challenging with more complex scenes, so a stronger TNE may lead to better localization robustness. Further, we show that directly applying the localization map to extract sounding visual features (Soft in Table S6) also achieves reasonable performances, it indicates that TNE does not rely on a hard threshold.

3. More Qualitative Results

To better understand the superiority of the learned model, we report more visualization results in Fig. S3, S4 and S5. As shown, our model can better localize the sound-source objects/regions in various audio-visual scenes.

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

[2] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725. IEEE, 2020. 1

[3] Di Hu, Feiping Nie, and Xuelong Li. Deep multimodal clustering for unsupervised audiovisual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9248–9257, 2019. 2

[4] Shentong Mo and Pedro Morgado. Localizing visual sounds the easy way. *arXiv preprint arXiv:2203.09324*, 2022. 2

[5] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 1, 2

[6] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*, 2019. 3

[7] Rui Qian, Di Hu, Heinrich Dinkel, Mengyue Wu, Ning Xu, and Weiyao Lin. Multiple sound sources localization from coarse to fine. In *European Conference on Computer Vision*, pages 292–308. Springer, 2020. 2

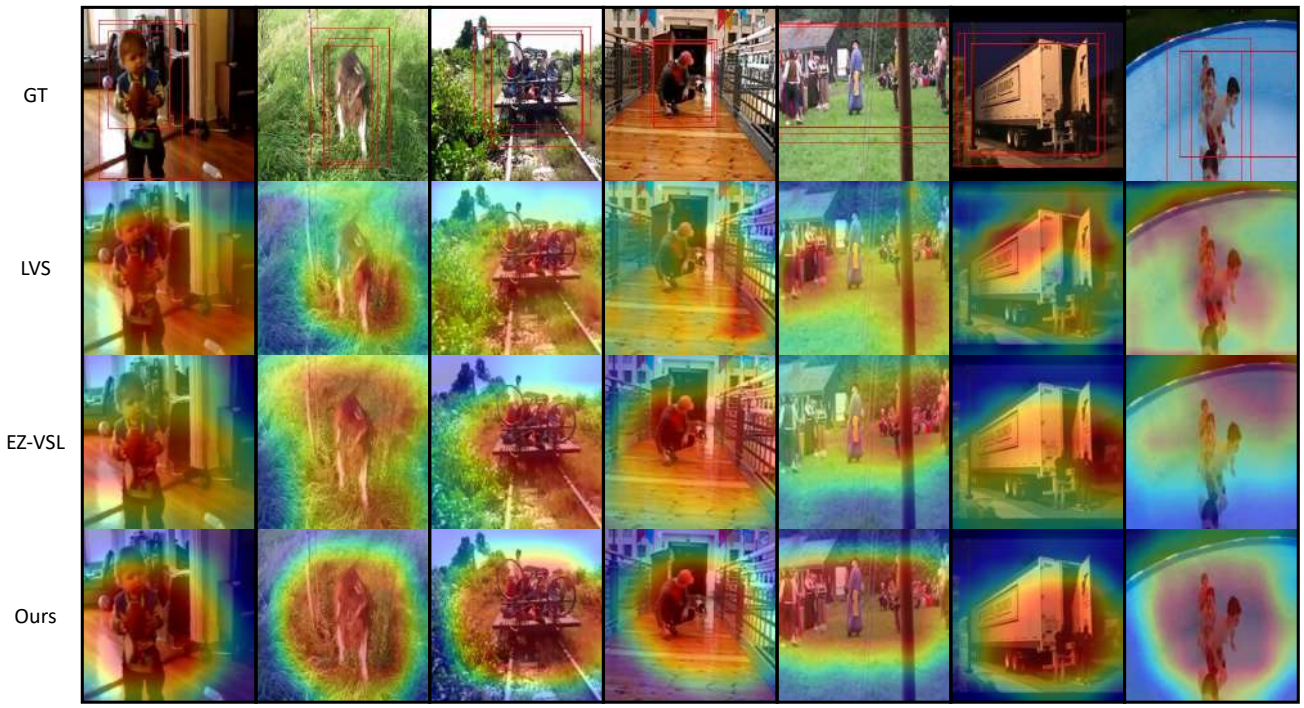

Figure S3. Qualitative results of FlickrSoundNet testset.

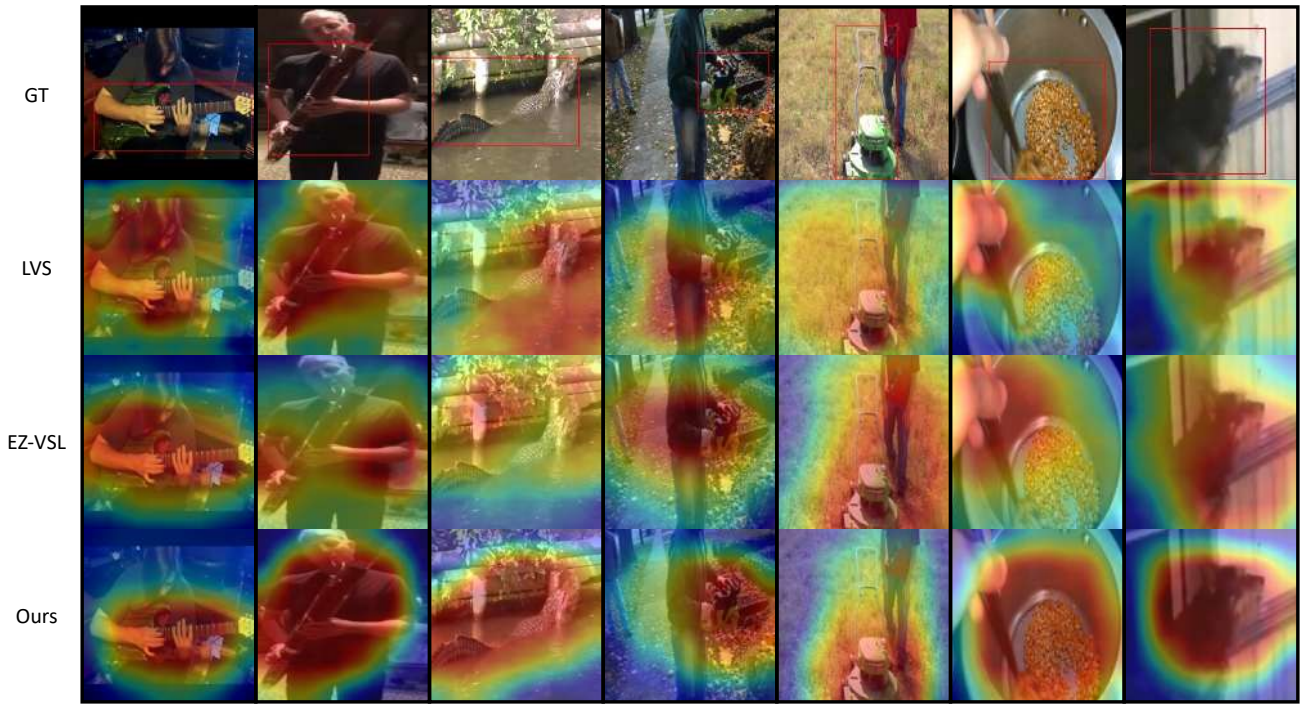

Figure S4. Qualitative results of VGGSound-Source.

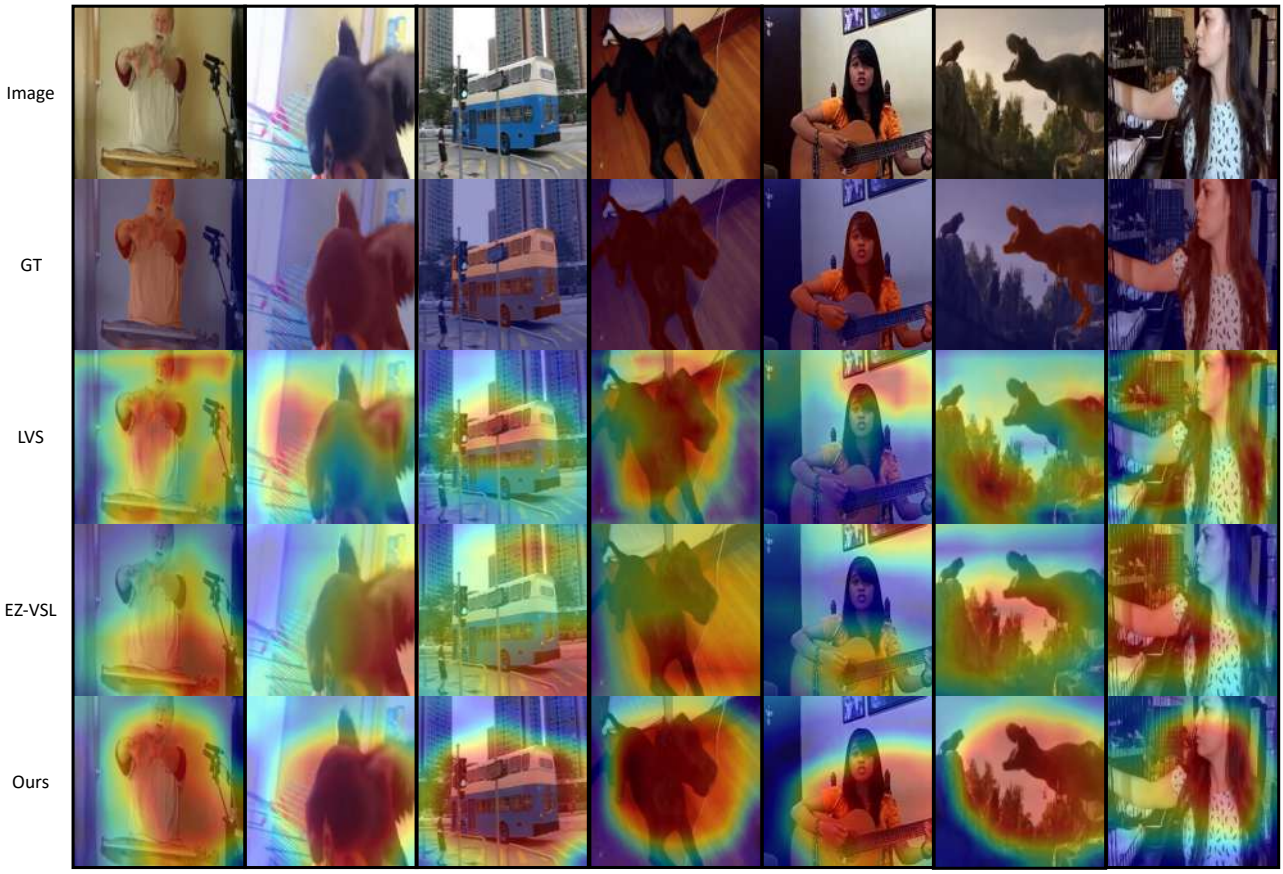

Figure S5. Qualitative results of AVSBench.