Sound Source Localization is All about Cross-Modal Alignment

Arda Senocak[1*] Hyeonggon Ryu[1*] Junsik Kim[2*] Tae-Hyun Oh[3,4]
Hanspeter Pfister[2] Joon Son Chung[1]
[1] KAIST [2] Harvard University [3]POSTECH [4]Yonsei University

Abstract

Humans can easily perceive the direction of sound sources in a visual scene, termed sound source localization. Recent studies on learning-based sound source localization have mainly explored the problem from a localization perspective. However, prior arts and existing benchmarks do not account for a more important aspect of the problem, cross-modal semantic understanding, which is essential for genuine sound source localization. Cross-modal semantic understanding is important in understanding semantically mismatched audio-visual events, e.g., silent objects, or off-screen sounds. To account for this, we propose a cross-modal alignment task as a joint task with sound source localization to better learn the interaction between audio and visual modalities. Thereby, we achieve high localization performance with strong cross-modal semantic understanding. Our method outperforms the state-of-the-art approaches in both sound source localization and cross-modal retrieval. Our work suggests that jointly tackling both tasks is necessary to conquer genuine sound source localization.

1. Introduction

Humans can easily perceive where the sound comes from in a scene. We naturally attend to the sounding direction and associate incoming audio-visual signals to understand the event. To achieve human-level audio-visual perception, sound source localization in visual scenes has been extensively studied [20, 21, 1, 19, 2, 13, 11, 12, 23, 24, 22, 14, 16, 15, 8]. Motivated by that humans learn from natural audio-visual correspondences without explicit supervision, most of the studies have been developed on a fundamental assumption that audio and visual signals are temporally correlated. With the assumption, losses of the sound source localization task are modeled by audio-visual correspondence as a self-supervision signal and are implemented by contrasting audio-visual pairs, *i.e.*, contrastive learning.

While these approaches appear to be unsupervised methods, they strongly rely on partial supervision information; *e.g.*, using supervisedly pretrained vision networks [20, 21, 19, 23, 24, 8] and visual objectness estimators for post-processing [16, 15]. Without leveraging such strong representations, the performance is degraded. Thus, the previous methods are not purely self-supervised approaches. Even further, there are recent studies [18, 16, 15] that point out visual objectness bias in existing sound source localization benchmarks and exploit the objectness prior to improve the localization accuracy. They show that, even without interaction between visual and audio signals, a model achieve strong accuracy in localization by referring visual signals alone, which is not the true intention of the sound source localization task. In short, the current evaluation do not capture the true sound source localization performance.

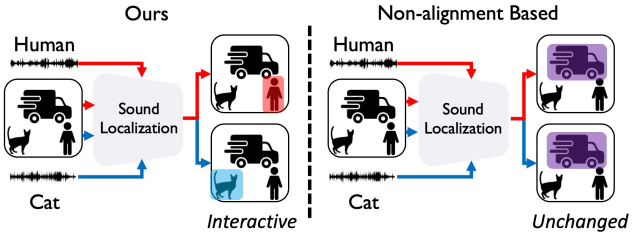

Figure 1. **A conceptual difference between prior approaches and our alignment-based sound source localization.**

In this work, we first sort out evaluating sound source localization methods by introducing a cross-modal retrieval task as an auxiliary evaluation task. By this task, we can measure whether the learned representation have the capability to accurately interact between audio and visual modalities; *i.e.*, more fine-grained audio-visual correspondence which is essential for genuine sound source localization. This aspect has been missed in existing sound source localization benchmarks. Indeed, our experiments show that higher sound localization performance does not guarantee higher cross-modal retrieval performance.

Second, given this additional criterion, we revisit the importance of semantic understanding shared across audio and visual modalities in both sound source localization and cross-modal retrieval. In the previous methods [20, 21, 24, 19], the cross-modal semantic alignment

*These authors contributed equally to this work.

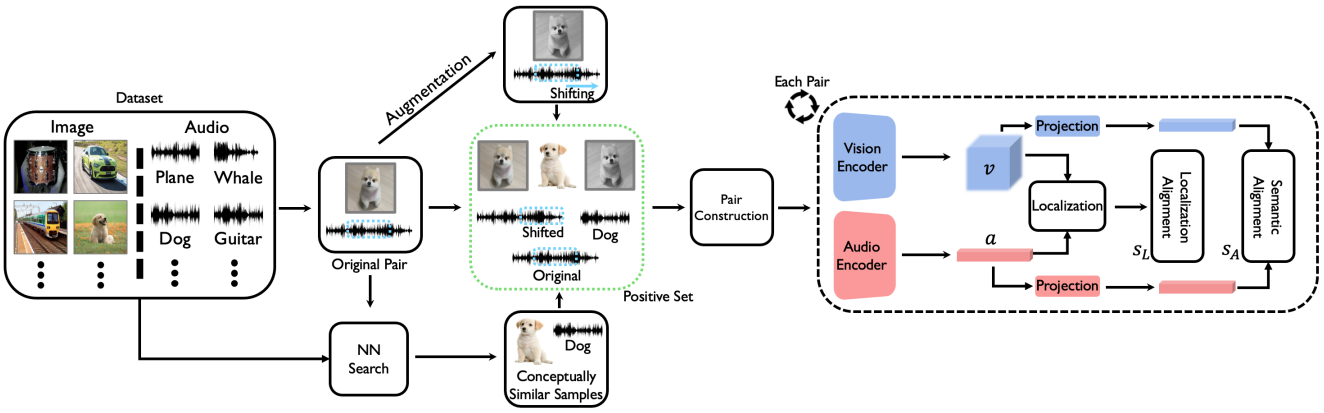

Figure 2. **Our sound source localization framework.** Our model construct multiple positive pairs with augmentation and Nearest Neighbor Search (Conceptually Similar Samples). By using these newly constructed 9 pairs, our model employs spatial localization, s_L, and semantic feature alignment, s_A, for each pair to learn a better sound source localization ability.

is induced by instance-level cross-modal contrastive learning, *i.e.*, cross-modal instance discrimination between visual and audio features. However, they are aided by labels or supervisedly pretrained encoder [2] for easing challenging cross-modal feature alignment. Instead, our method learns from scratch supporting the lack of guidance by incorporating multiple positive samples into cross-modal contrastive learning. Specifically, we construct a positive set for each modality using both multi-view [3] and conceptually similar samples [7]. Thereby, we enhance feature alignment and achieve high localization performance and strong cross-modal semantic understanding.

We evaluate our method on the VGG-SS and SoundNet-Flickr benchmarks for sound source localization. As aforementioned, the sound source localization task is closely related to the cross-modal retrieval task, but our experiments show that existing works have a weak performance correlation between them. This implies that we need to evaluate both tasks for evaluating the genuine sound source localization. The proposed method performs favorably against the recent state-of-the-art approaches in both tasks.

2. Related work

Sound source localization in visual scenes has been investigated by exploiting correspondences between audio and visual modalities. The most widely used approach for sound source localization is cross-modal attention [20, 21, 25] with contrastive loss [5, 10, 17]. Later, the attention-based method is improved by intra-frame hard sample mining [2], iterative contrastive learning with pseudo labels [13], feature regularization [14], positive mining [22], negative free learning [24] with stop-gradient operation [4], or momentum encoders [15]. Some sound localization approaches exploit additional semantic labels [19, 12, 23] or object prior [16, 26]. Semantic labels are used to pretrain audio and vision encoders with classification loss [12, 23] or refine audio-visual feature alignment [19]. A more explicit way to refine localization output is to use object prior. EZVSL [16] proposes post-processing to combine attention based localization output with a pretrained visual feature activation map. However, postprocessing by object prior may generate a false positive output as it is solely based on vision without audio-visual interaction.

[2]Typically, an image encoder is pretrained on ImageNet [6] and an audio encoder is pretrained on AudioSet [9] in supervised ways.

3. Method

3.1. Cross-Modal Feature Alignment

We consider both spatial localization and semantic feature alignment for sound source localization. To this end, we use two different similarity functions s_L and s_A for contrastive learning (more specifically InfoNCE), s_L for localization and s_A for cross-modal feature alignment.

Recent studies rely on audio-visual spatial correspondence maps to learn sound source localization by contrasting them. Given a spatial visual feature $\mathbf{v} \in \mathbb{R}^{c \times h \times w}$ and audio feature $\mathbf{a} \in \mathbb{R}^{\mathbf{c}}$, audio-visual similarity with a correspondence map can be calculated as follows:

$$s_L(\mathbf{v}, \mathbf{a}) = \sum_{xy \in M} \frac{1}{|M|} \frac{\mathbf{v}^{xy} \cdot \mathbf{a}}{\|\mathbf{v}^{xy}\| \|\mathbf{a}\|} \tag{1}$$

where $\mathbf{v}^{xy}$ is a feature vector at location (x, y), and M is an optional binary mask when an annotation or pseudo-mask [2, 14] is available. Since we assume no supervision for sound source localization, we do not use any mask, therefore, $M = \mathbf{1}$.

The contrastive loss with localization similarity s_L enforces location dependent alignment giving sparse but strong audio-visual correspondence which enables to perform localization. However, our empirical studies on cross-modal retrieval indicate that strong localization performance does not guarantee semantic understanding. To overcome the low semantic understanding in recent studies, we

propose to add instance-level contrastive loss. Instance-level contrasting encapsulates the whole context in a scene, enforcing better audio-visual semantic alignment. However, instance-level contrasting may smooth out spatial discriminativeness learned by Eq. (1). Inspired by SimCLR [3], we adopt a projection layer to align audio-visual semantics in a projection space. The projection layer separates the latent space of localization and semantic alignment, thereby preventing the alignment loss smoothing out the spatial discriminativeness. The similarity function for cross-modal feature alignment is defined as follows:

$$s_A(\mathbf{v}, \mathbf{a}) = \frac{p_v(\mathsf{avg}(\mathbf{v})) \cdot p_a(\mathbf{a})}{\|p_v(\mathsf{avg}(\mathbf{v}))\| \|p_a \mathbf{a}\|} \tag{2}$$

where $\mathsf{avg}(\cdot)$ is spatial average pooling, p_v is a projection layer for visual features, and p_a is a projection layer for audio features.

3.2. Expanding with Multiple Positive Samples

Typically, contrastive learning contrasts between one positive pair and multiple negative pairs. In audio-visual learning, by an audio-visual correspondence assumption, an audio-image pair from the same clip is used as a positive pair while negative pairs are sampled from different clips. However, single-instance discrimination may not be sufficient to achieve strong cross-modal alignment. We expand contrastive learning beyond single instance discrimination by positive set construction and pairing them. To construct a positive set, we incorporate both hand-crafted positive and conceptual positive samples for each modality. Later, we adjust the contrastive learning to incorporate multiple positive pairs to enforce cross-modal alignment.

Obtaining hand-crafted positive samples. Using randomly augmented samples as positive multi-view pairs are widely adopted in self-supervised representation learning. Similarly, we extend a single anchor audio-image pair to multiple positive pairs by applying simple augmentations on image and audio samples separately. While we utilize common image transformations on images, we apply temporal shifting to audios. It is worth noting that sound source localization task learns from the semantic consistency rather than subtle time differences as in videos. Thus, a slight shift in the audio may not alter contextual information significantly. As a result of hand-crafted multi-view positive pair generation, we obtain additional $\mathbf{v}^{aug}$ and $\mathbf{a}^{aug}$ samples.

Obtaining conceptual positive samples. Apart from manually created augmented views, we expand our positive set with conceptually similar samples. For selecting similar samples, we utilize pretrained encoders. Note that pretrained encoders trained either with supervised or self-supervised learning are effective in positive sample mining. By employing readily available image and audio encoders, we use the k-nearest neighborhood search to sample semantically similar samples in both modalities. In particular, given a pair of image and audio, we compute cosine similarity with all other samples and choose the top-k most similar samples among the training set for each modality. From a set of k samples, we randomly select one sample to obtain conceptually similar samples for each modality, $\mathbf{v}^{conc.}$ and $\mathbf{a}^{conc.}$. By utilizing the similar samples as positive samples, our model expands semantic understanding.

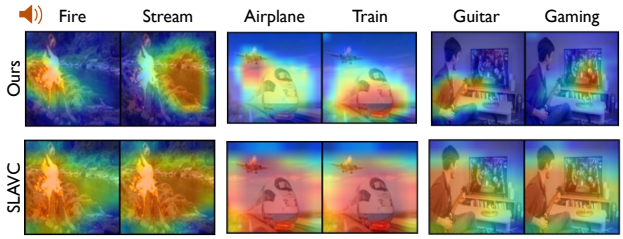

Figure 3. **Interactive Sound Localization of Ours and SLAVC [15].** Our model correctly follows the cross-modal interaction for given different sounds.

Pair Construction. Once we obtain the conceptual and hand-crafted positive samples for each modality, we create 9 distinct audio-visual pairs by pairing $\mathbf{V} = \{\mathbf{v}, \mathbf{v}^{aug}, \mathbf{v}^{conc}\}$ and $\mathbf{A} = \{\mathbf{a}, \mathbf{a}^{aug}, \mathbf{a}^{conc}\}$. This is done to ensure semantic alignment and consistency between them through contrastive learning. The negative pairs are randomly paired from the remaining samples in a training set. Note that some of these pairs are a combination of hand-crafted and conceptually similar samples, which further enhances the feature alignment of our model during training.

3.3. Training

Our loss formulation incorporates both localization and instance-level similarity functions with multiple positive pairs constructed by augmentation and conceptually similar sample search. The final loss term is defined as follows:

$$\mathcal{L}_i = -\sum_{\mathbf{v}_i \in \mathbf{V}} \sum_{\mathbf{a}_i \in \mathbf{A}} \left[\log \frac{\exp(s_L(\mathbf{v}_i, \mathbf{a}_i)/\tau)}{\sum_j \exp(s_L(\mathbf{v}_i, \mathbf{a}_j)/\tau)} + \log \frac{\exp(s_A(\mathbf{v}_i, \mathbf{a}_i)/\tau)}{\sum_j \exp(s_A(\mathbf{v}_i, \mathbf{a}_j)/\tau)} \right] \tag{3}$$

where $\mathbf{V}$ and $\mathbf{A}$ indicate positive sample sets.

4. Experiment Results

Quantitative comparison with strong baselines. In this section, we conduct a comparative analysis of our sound source localization method against existing approaches. We train our model on VGGSound-144K and evaluate it on VGG-SS and SoundNet-Flickr test sets. We present our results in Table 1. Our proposed model achieves higher performance compared to prior approaches on both test sets.

Method	Pre. Vision	VGG-SS cIoU ↑	VGG-SS AUC ↑	Flickr-SoundNet cIoU ↑	Flickr-SoundNet AUC ↑
Attention [20]$_{CVPR18}$	✓	18.50	30.20	66.00	55.80
CoarseToFine [19]$_{ECCV20}$	✓	29.10	34.80	-	-
LCBM [23]$_{WACV22}$	✓	32.20	36.60	-	-
LVS [2]†$_{CVPR21}$	✗	30.30	36.40	72.40	57.80
LVS [2]$_{CVPR21}$	✗	34.40	38.20	71.90	58.20
HardPos [22]$_{ICASSP22}$	✗	34.60	38.00	76.80	59.20
SSPL (w/o PCM) [24]$_{CVPR22}$	✓	27.00	34.80	73.90	60.20
SSPL (w/ PCM) [24]$_{CVPR22}$	✓	33.90	38.00	76.70	60.50
EZ-VSL (w/o OGL) [16]$_{ECCV22}$	✓	35.96	38.20	78.31	61.74
SSL-TIE [14]$_{ACM\ MM22}$	✗	38.63	39.65	79.50	61.20
SLAVC (w/o OGL) [15]$_{NeurIPS22}$	✓	37.79	39.40	**83.60**	-
Ours					
↳ NN Search w/ Supervised Pre. Encoders	✗	**39.94**	**40.02**	79.60	**63.44**
↳ NN Search w/ Self-Supervised Pre. Encoders	✗	39.20	39.70	79.20	63.00
with OGL:					
EZ-VSL (w/ OGL) [16]$_{ECCV22}$	✓	38.85	39.54	83.94	63.60
SLAVC (w/ OGL) [15]$_{NeurIPS22}$	✓	39.80	-	**86.00**	-
Ours (w/ OGL)					
↳ NN Search w/ Supervised Pre. Encoders	✗	**42.64**	**41.48**	82.40	**64.60**
↳ NN Search w/ Self-Supervised Pre. Encoders	✗	42.47	41.42	82.80	64.48
with Optical Flow:					
HearTheFlow [8]$_{WACV23}$	✓	39.40	40.00	84.80	64.00

Table 1. **Quantitative results on the VGG-SS and SoundNet-Flickr test sets**. All models are trained with 144K samples from VGG-Sound and tested on VGG-SS and SoundNet-Flickr.

Specifically, it yields a +2.15% cIoU and +0.6% AUC improvement on VGGSS, as well as a +3.7% cIoU improvement on SoundNet-Flickr compared to the state-of-the-art methods that uses pretrained vision encoder. It is worth highlighting that unlike the majority of previous works, our proposed model does not utilize a vision encoder pretrained on ImageNet in a sound source localization backbone. This is because, as discussed in Mo *et al*. [15], using supervisedly pretrained vision encoders makes the sound source localization problem a weakly supervised problem. However, it is worth noting that even without using a pretrained vision encoder, our method achieves state-of-the-art performance on both experiments that are presented in Table 1.

We demonstrate the performance of our model with the pretrained models learned through supervised learning (NN Search w/ Supervised Pre. Encoders) and with models that are pretrained through self-supervised learning (NN Search w/ Self-Supervised Pre. Encoders) in NN Search module. As the results indicate, using self-supervised pretrained encoders in NN Search performs on par with the supervised pretrained encoders in NN Search. This shows that our model can utilize any type of pretrained encoder feature for nearest neighbor search. Note that these pretrained encoders are not used in the backbone networks of the sound source localization module but only in the NN Search Module, as illustrated in Figure 2. Our model do not require any task-specific modules or operations to achieve the state-of-the-art results. This suggests that using additional semantic and multi-view correspondence, as well as feature alignment, provides more varied and robust supervision for better aligned audio and visual features, as opposed to using task-specific approaches.

Retrieval. We evaluate sound localization models on the VGG-SS dataset for cross-modal retrieval. As shown in Table 2, our method clearly outperforms other state-of-the-art

Model	Pre. Vision	A → I R@1	A → I R@5	A → I R@10	I → A R@1	I → A R@5	I → A R@10
LVS [2]$_{CVPR21}$	✗	3.87	12.35	20.73	4.90	14.29	21.37
EZ-VSL [16]$_{ECCV22}$	✓	5.01	15.73	24.81	14.2	33.51	45.18
SSL-TIE [14]$_{MM22}$	✗	10.29	30.68	43.76	12.76	29.58	39.72
SLAVC [15]$_{NeurIPS22}$	✓	4.77	13.08	19.10	6.12	21.16	32.12
Ours							
↳ NN Search w/ Supervised Pre. Encoders	✗	**16.47**	36.99	49.00	**20.09**	**42.38**	**53.66**
↳ NN Search w/ Self-Supervised Pre. Encoders	✗	14.31	**37.81**	**49.17**	18.00	38.39	49.02

Table 2. **Summary of retrieval recall scores for all models.** All of the models are trained on VGGSound 144K data and retrieval is performed on entire VGG-SS dataset, containing ∼5K samples.

methods. One interesting observation is that EZ-VSL [16] notably performs better than SLAVC [15] on cross-modal retrieval, while SLAVC performs better on sound source localization in Table 1. This shows that with the current benchmark evaluations, better sound localization performance does not guarantee better audio-visual semantic understanding, thereby we need to additionally evaluate sound source localization methods on cross-modal understanding tasks. Another observation is that the performance gap between our method and the strongest competitor SSL-TIE [14] is notably larger on cross-modal retrieval than sound source localization. This is due to the strong cross-modal feature alignment of our method that is overlooked in the sound source localization benchmarks.

Qualitative Results. We visualize and compare our sound localization results with the recent prior work. We demonstrate interactiveness of our method across modalities in Figure 3. Genuine sound source localization should be able to localize objects that are correlated with the sound. To visualize cross-modal interaction, we synthetically pair the same image with different sounds of objects that are visible in a scene. The examples demonstrate that the proposed method can localize different objects depending on the contexts of sounds, while the competing method can not.

5. Conclusion

In this work, we investigate cross-modal semantic understanding that has been overlooked in sound source localization studies. We observe that higher sound source localization performance on the current benchmark does not necessarily show higher performance in cross-modal retrieval, despite its causal relevance in reality. To enforce strong understanding of audio-visual semantic matching while maintaining localization capability, we propose semantic alignment with multi-views of audio-visual pairs in a simple yet effective way. The ablation study shows that strong semantic alignment is achieved when both semantic alignment loss and enriched positive pairs are used. We extensively evaluate our method on sound source localization benchmarks. Moreover, our analyses on cross-modal retrieval and false positive detection verify that the proposed method has strong capability in cross-modal interaction. Our study suggests that sound localization methods should be evaluated not only on localization benchmarks but also on cross-modal understanding tasks.

References

[1] Relja Arandjelović and Andrew Zisserman. Objects that sound. In *European Conference on Computer Vision*, 2018. 1

[2] Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Localizing visual sounds the hard way. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 1, 2, 4

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, 2020. 2, 3

[4] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 2

[5] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2005. 2

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009. 2

[7] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. In *IEEE International Conference on Computer Vision*, 2021. 2

[8] Dennis Fedorishin, Deen Dayal Mohan, Bhavin Jawade, Srirangaraj Setlur, and Venu Govindaraju. Hear the flow: Optical flow-based self-supervised visual sound source localization. In *IEEE Winter Conf. on Applications of Computer Vision (WACV)*, 2023. 1, 4

[9] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2017. 2

[10] Elad Hoffer and Nir Ailon. Deep metric learning using triplet network. In *International workshop on similarity-based pattern recognition*, 2015. 2

[11] 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, NeurIPS*, 2020. 1

[12] Sizhe Li, Yapeng Tian, and Chenliang Xu. Space-time memory network for sounding object localization in videos. In *British Machine Vision Conference*, 2021. 1, 2

[13] Yan-Bo Lin, Hung-Yu Tseng, Hsin-Ying Lee, Yen-Yu Lin, and Ming-Hsuan Yang. Unsupervised sound localization via iterative contrastive learning. *arXiv preprint arXiv:2104.00315*, 2021. 1, 2

[14] Jinxiang Liu, Chen Ju, Weidi Xie, and Ya Zhang. Exploiting transformation invariance and equivariance for self-supervised sound localisation. In *ACM International Conference on Multimedia*, 2022. 1, 2, 4

[15] Shentong Mo and Pedro Morgado. A closer look at weakly-supervised audio-visual source localization. *Advances in Neural Information Processing Systems, NeurIPS*, 2022. 1, 2, 3, 4

[16] Shentong Mo and Pedro Morgado. Localizing visual sounds the easy way. In *European Conference on Computer Vision*, 2022. 1, 2, 4

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

[18] Takashi Oya, Shohei Iwase, Ryota Natsume, Takahiro Itazuri, Shugo Yamaguchi, and Shigeo Morishima. Do we need sound for sound source localization? In *Asia Conference on Computer Vision*, 2020. 1

[19] 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*, 2020. 1, 2, 4

[20] 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. 1, 2, 4

[21] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound sources in visual scenes: Analysis and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(5):1605–1619, 2021. 1, 2

[22] Arda Senocak, Hyeonggon Ryu, Junsik Kim, and In So Kweon. Learning sound localization better from semantically similar samples. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2022. 1, 2, 4

[23] Arda Senocak, Hyeonggon Ryu, Junsik Kim, and In So Kweon. Less can be more: Sound source localization with a classification model. In *IEEE Winter Conf. on Applications of Computer Vision (WACV)*, 2022. 1, 2, 4

[24] Zengjie Song, Yuxi Wang, Junsong Fan, Tieniu Tan, and Zhaoxiang Zhang. Self-supervised predictive learning: A negative-free method for sound source localization in visual scenes. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 1, 2, 4

[25] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *European Conference on Computer Vision*, 2018. 2

[26] Hanyu Xuan, Zhiliang Wu, Jian Yang, Yan Yan, and Xavier Alameda-Pineda. A proposal-based paradigm for self-supervised sound source localization in videos. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1029–1038, 2022. 2