Learning from Silence and Noise for Visual Sound Source Localization

Xavier Juanola[1]
xavier.juanola@upf.edu
Giovana Morais[2]
giovana.morais@nyu.edu
Magdalena Fuentes[2]
mfuentes@nyu.edu
Gloria Haro[1]
gloria.haro@upf.edu

[1] Intelligent Multimodal Vision Analysis
Universitat Pompeu Fabra
Barcelona, Spain

[2] MARL-IDM
New York University,
New York, USA

Abstract

Visual sound source localization is a fundamental perception task that aims to detect the location of sounding sources in a video given its audio. Despite recent progress, we identify two shortcomings in current methods: 1) most approaches perform poorly in cases with low audio-visual semantic correspondence such as silence, noise, and offscreen sounds, i.e. in the presence of negative audio; and 2) most prior evaluations are limited to positive cases, where both datasets and metrics convey scenarios with a single visible sound source in the scene. To address this, we introduce three key contributions. First, we propose a new training strategy that incorporates silence and noise, which improves performance in positive cases, while being more robust against negative sounds. Our resulting self-supervised model, SSL-SaN, achieves state-of-the-art performance compared to other self-supervised models, both in sound localization and cross-modal retrieval. Second, we propose a new metric that quantifies the trade-off between alignment and separability of auditory and visual features across positive and negative audio-visual pairs. Third, we present IS3+, an extended and improved version of the IS3 synthetic dataset with negative audio. Our data, metrics and code are available on the Project page.

1 Introduction

Humans have an incredible capacity for multimodal integration, particularly in processing audio-visual information from the environment. While sound localization is primarily handled by the auditory system, visual data aids in disambiguation and accuracy [51]. Inspired by human perceptual abilities, multimodal approaches have demonstrated that combining auditory and visual information not only improves traditional perception tasks, such as object detection and activity recognition [9, 13, 53], but also enables entirely new capabilities. For example, audio-visual scene synthesis allows systems to reconstruct missing modalities [26, 43, 67], predict soundscapes for silent videos [12], or even anticipate visual scenes based on sound [34, 59, 63]. Other examples that benefited from multimodality are audio-visual

source separation [2, 16, 42], comprehensive scene understanding [14, 32] and visual sound source localization.

Visual sound source localization (VSSL) aims to localize sounding sources within a video. Pioneering approaches used networks pretrained in ImageNet [11] as backbones and finetuned them for audio-visual correspondence [6, 39, 40], resulting in weakly-supervised models [40]. Recently, new methods, such as [35, 58], shifted from weakly-supervised learning to self-supervised learning, training models from scratch and using different data augmentation techniques, *e.g*. geometrical transformations, to avoid overfitting and improve robustness. However, these augmentations are mostly done in positive cases (*i.e*. when there is a sounding source that is *visible* in the image), and thus these models still struggle when presented with negative cases [6, 27].

Moreover, current datasets used to benchmark VSSL methods, such as VGGSound [5], VGG Sound Sources (VGG-SS) [6], and Flickr [3], feature mostly sounding objects that are in the foreground and centered in the image frame [56]. As a result, there is no strong incentive for models to use the audio content to localize the sound source in the image. Instead, models tend to identify the object regardless of sound [46]. Recent efforts have been made to mitigate this by introducing synthetic evaluation sets with multiple sources [58], but these still suffer from inaccurate image-audio pairings, and model evaluation is mostly done on positive cases.

To address these shortcomings, we introduce three main contributions: (1) A novel training strategy that incorporates negative audio samples (silence and noise) during training, with two additional loss terms that ensure the network learns to ignore these non-informative sounds; (2) The IS3+ benchmark, an improved version of IS3 [58], in which we substitute incorrect image-audio pairs that were present in the original test set with correct audio from Adobe Sound Effects [1] and a clean subset of IS3 audio samples. In addition, we introduce the evaluation of negative audio cases, namely in the presence of noise, silence, and offscreen sounds; and (3) A new metric that quantifies the discriminative power (or separability) of auditory and visual features, which correlates with performance in both sound localization and cross-modal retrieval tasks. Through a comprehensive evaluation of sound localization models in both positive and negative audio samples, we show that our model SSL-SaN (Sound Source Localization with Silence and Noise), trained with our new strategy, yields state-of-the-art results. Code and data are available at https://xavijuanola.github.io/SSL-SaN/.

2 Related Work

Pioneering works [15, 23, 28] learn to capture correspondences between audio and visual features using classical machine learning methods, such as canonical correlation analysis [20]. More recent methods adopted deep neural networks for representation learning by leveraging the synchronization between audio and video as a signal for self-supervised learning [31, 45]. Currently, the most widely used approach for sound source localization is cross-modal attention [52, 55, 64] with contrastive loss [6, 10, 44, 45, 54]. These approaches seek to localize objects by aligning audio and visual representation spaces. Some methods use additional semantic labels to pretrain audio and vision with classification loss [57] or refine audio-visual feature alignment [49]. Knowledge distillation from pretrained object detection and sound classification models was used in [67]. LVS [6] used a contrastive loss with hard negative mining to learn the audio-visual co-occurrence map discriminatively. EZ-VSL [39] introduced a multiple instance contrastive learning framework that focuses only on the most aligned regions when matching the audio to the video by combining the attention-based

localization output with a pretrained visual feature activation map. ACL [48] leverages the CLIPSeg model [37], which is an image segmentation model trained in a supervised way.

Most works in the literature were trained and evaluated with datasets that in majority feature a single-sounding object present in the scene at a given time [3, 4, 6, 35, 39, 40, 45, 50, 54, 58, 60, 61, 66]. This setting is rare in real-life scenarios, where there are multiple objects sounding at the same time (*i.e.* a mixture of sounds), silent objects, sounds produced by objects that are not visually present in the scene or occluded by other objects (offscreen sounds), noise, etc. There has been an increasing interest in working with mixtures of sounds [25, 29, 38, 41, 49], but very few worked with silent objects [24, 36, 40] or offscreen sounds [36]. DSOL [24] and IEr [36] proposed a method to suppress localization of silent objects by creating an Audio-Instance-Identifier module, which identifies the sounds present in the audio, and filters out possible offscreen sounds and silent objects. These methods, however, rely on the number of sound sources, which is not available in most large-scale datasets, or "in-the-wild" data. Given the complexity of the problem and the limitations of the datasets available for training, many models in this domain are prone to overfitting. As a result, early stopping is commonly adopted as a practical regularization technique to prevent overfitting and improve generalization performance on unseen data. SLAVC [40], on the other hand, proposes a framework that solves overfitting and the need for early stopping by adding extreme visual dropout and momentum encoders. Another key strategy to overcome data limitations is data augmentation [7, 8, 18, 22]. Following this, SSL-TIE [35] presents a neural network composed of an image and an audio encoder, trained with contrastive learning and geometrical consistency, ensuring that the audio-visual similarity maps undergo the same geometrical transformation as the input images. SSL-Align [58] proposes a novel method that utilizes semantic alignment with multi-views and semantically similar samples. The authors present both a fully self-supervised and a weakly-supervised variation –with supervisedly pretrained audio and image encoders– of their model.

Except for a few works [19, 27, 40], most evaluations are predominantly done on positive cases (*i.e.* visible *and* audible sources). This not only results in incomplete assessments of the model capabilities, but also reinforces biases in models, as they are implicitly optimized and judged under conditions that favor co-occurring signals. To address this, we combine image and audio augmentations with additional negative audio samples (silence and noise) and additional loss terms during training, resulting in a more robust and accurate self-supervised model. This fully self-supervised model outperforms previous self-supervised VSSL models, as we demonstrate in a comprehensive evaluation including both positive and negative cases. In a concurrent work [33], and in a different context –supervised audio-visual segmentation– the inclusion of silence and noise during training has also been shown to be beneficial.

3 Method

Most of the VSSL models (*e.g.* [4, 6, 24, 35, 39, 47, 56, 57, 58, 61, 66]) use a two-stream network with an audio encoder that extracts auditory features, $\mathbf{a}_i \in \mathbb{R}^c$, from the i-th audio segment and a visual encoder that extracts visual features, $\mathbf{v}_j \in \mathbb{R}^{c\times h\times w}$, from the j-th image (usually the central frame of the j-th audio segment). Then, an audio-visual similarity map is computed by cosine similarity:

$$S(\mathbf{a}_i, \mathbf{v}_j) = \frac{\mathbf{a}_i \cdot \mathbf{v}_j}{\|\mathbf{a}_i\| \cdot \|\mathbf{v}_j\|} \in [-1,1]^{h\times w}, \tag{1}$$

and a global audio-visual correspondence value is computed from S by a certain pooling operation [9, 6, 24, 35, 39, 47] and eventually semantic projection heads [58]. The network is trained in a self-supervised way by contrastive learning; forming positive and negative audio-visual pairs just by taking $i = j$ and $i \neq j$, respectively.

Typically, *e.g.* [6, 24, 35, 39, 40, 47, 56, 57, 61], the audio and visual encoders are ResNet18 networks. The audio encoder is trained from scratch and the visual encoder is initialized with ImageNet pretrained weights, resulting in weakly-supervised models [40]. Recent works, [35, 58], have shown that state-of-the-art results in VSSL can be achieved by training both encoders from scratch (*i.e.* in a fully self-supervised way). Both of them use (different) data augmentation techniques. Additionally, [35] uses an equivariance loss term and [58] uses a semantic projection head on top of the encoders. As baseline model we use SSL-TIE [35], since it is a state-of-the-art self-supervised model [27] whose code is open-source. However, our proposed training strategy can be applied to any localization model trained in a contrastive way.

3.1 Learning from Silence and Noise

Our aim is to design a sound localization model that is robust to negative sounds such as silence and noise. To that end, we introduce two modifications during training. First, we pair each image in the batch with these two types of negative audio samples, thus creating two new negative audio-visual pairs. We define *silence* as an empty audio and *noise* as an audio with random values following a Gaussian distribution with zero mean and standard deviation σ=1. Second, we add two new loss terms forcing an empty similarity map for these two negative audio-visual pairs. More concretely, for every j-th image in the batch we define:

$$\mathcal{L}_S = \left\| S(\mathbf{a}^S, \mathbf{v}_j) \right\|_2^2,$$

the square L_2 norm of the audio-visual similarity map between the visual features from the j-th image, $\mathbf{v}_j$, and the silence feature, $\mathbf{a}^S$, and analogously, for the same visual features and the audio features extracted from a realization of the noise, $\mathbf{a}_j^N$, that is:

$$\mathcal{L}_N = \left\| S(\mathbf{a}_j^N, \mathbf{v}_j) \right\|_2^2.$$

Intuitively, these terms penalize the model localizing any sound in the presence of silence and noise, enforcing a predictable behavior in the presence of negative audio. Moreover, as shown in the experimental section, learning from silence and noise also improves the results with a positive audio.

3.2 New evaluation set: IS3+

The current VSSL benchmarks, such as VGG-SS [6], contain videos gathered from YouTube, where typically a single object dominates the scene, both in the auditory and visual modalities. Consequently, the task of sound localization can be solved by using objectness cues in the image, without the need for the audio characteristics [46]. This fact motivated the proposal, in [58], of a new VSSL benchmark: *Interactive-Synthetic Sound Source (IS3)*. IS3 spans 118 object categories and includes 3,240 images that have been synthetically generated by diffusion models [52], simulating scenes with multiple objects in diverse sizes. Each image in IS3 is paired with two audio samples from VGG-SS, corresponding to the two most visible

objects present in the image, resulting in 6,480 audio-visual instances. Although IS3 improves upon the existing benchmarks in terms of visual data, it still inherits the weaknesses of VGG-SS in the audio modality. VGG-SS videos are in-the-wild videos from YouTube and in a significant number of videos the audio signal does not contain the sound of the object in the scene. Typical examples of wrong audio samples are music instead of the original audio, an offscreen sound that masks the sound of interest, or a mixture of different types of sounds (see Supplementary material for more details). Thus, we propose an improved version of IS3, named IS3+, where each IS3 image is paired with two *clean* audio samples corresponding to the two main objects in the image.

To do so, we use a combination of audio samples from Adobe Sound Effects (Adobe SFX) [1], and clean samples from IS3. We manually selected samples from Adobe SFX through careful review and semantic matching of the audio content with IS3 categories. When we did not have enough Adobe SFX audio samples for a given class, we selected clean audio samples from IS3. The criteria for audio selection, in both Adobe SFX and IS3, were 1) to have audio that matches the necessary category, and 2) to avoid as much background noise as possible. We also simplified the dataset categories in which the image did not match the class. For example, the "playing harpsichord" category only had images of harps, therefore we replaced it with "harp". Similarly, we simplified categories such as "cat growling", "cat caterwauling" by "cat" as the images did not reflect these actions (the full mapping is in the Supplementary material). Finally, we create IS3+ by taking the simplified IS3 annotations and replacing the audio with a randomly sampled item within the correct category from the clean audio pool. We normalize and downsample audio samples to 16kHz.

4 Experiments and Results

4.1 Experimental Setup

Datasets. Our method is trained using the *VGGSound-144K* [54], a fixed subset of 144K videos randomly selected from VGGSound [5]. We test models' performance on *VGG Sound Sources* (*VGG-SS*) [4](a subset of VGG-Sound with annotated bounding boxes for sound sources), IS3 [58], IS3+ (presented in Section 3.2) and AVS-Bench S4 test set [69]. We evaluate the VGG-SS test set using the bounding boxes from the annotations, while we use the segmentation masks present in IS3, IS3$^+$ and AVS-Bench S4.

Evaluation Metrics. The current standard metric to evaluate sound source localization performance is the consensus Intersection over Union (cIoU) proposed by [54]. To evaluate the cIoU, a threshold is necessary to binarize the audio-visual similarity map (1) and convert it to a localization mask. Recently, [58] proposed the *Adaptive* cIoU, which sets the top B pixels to 1, where B is the area of the ground truth bounding box or mask. This metric assumes that the object is always visible and that the object's size is known, assumptions that do not hold in most cases. On the other hand, [27] proposed a *Universal threshold* that does not assume a visible object in the scene, and is designed to be discriminative of positive vs. negative audio cases. This threshold is set to be the 3rd quartile of the maximum audio-visual similarity in negative cases, so that it filters false positives without any assumption on the object size. We report cIoU with both the Universal threshold (Uth) and the adaptive one (Adap.). We also report the Area Under the Curve (AUC), which measures the integral of the success ratio (proportion of samples with cIoU$> \tau$) as a function of the threshold τ varying from 0 to 1. Besides these metrics, we report the percentage of Image Area (pIA) [27], which is designed to assess models' performance with negative audio by measuring the proportion of the image

area that has been activated in the model's localization mask in the presence of noise, silence and offscreen sounds separately. We also report AUC_N which is analogous to AUC for the case of negative audio (in this case, the success ratio is defined as proportion of samples with pIA $\leq \tau$). Finally, we evaluate overall model performance across both positive and negative cases with F_{LOC} *and* F_{AUC} [27]. They both compute the harmonic mean between a positive and a negative metric (F_{LOC} uses the cIoU and pIA while F_{AUC} uses AUC and AUC_N).

Implementation. Following [35], we resize input images into a resolution of 224×224, and we represent audio samples by log-Mel Spectrograms, extracted from 3 seconds of audio at a sample rate of 16 kHz. The log-Mel Spectrograms are computed using 512 as the size of the FFT windows, 239 for the hop length between STFT windows, and 257 mel filterbanks. We use ResNet18 [21] for the audio and image encoders. We train all models from scratch. The models are trained with a batch size of 32, the Adam optimizer [30] with a learning rate of $1e^{-4}$ and ReduceOnPlateau as the learning rate scheduler.

4.2 Comparison to prior work

We compare our model to different prior works: RCGrad [66], LVS [6], EZ-VSL [39], FNAC [61], SLAVC [40], ACL [48], SSL-Align [58], and SSL-TIE [35]. SSL-TIE and a version of SSL-Align are, as ours, the only fully sef-supervised models. The rest are weakly-supervised, since they leverage image encoders/decoders pretrained in a supervised way on annotated datasets such as ImageNet [11] or PhraseCut [65] (in case of ACL). We exclude object guided localization (OGL) postprocessing [39], as it is not meaningful for negative sounds.

Figure 1: Distribution of the maximum values of the audio-visual similarity maps across different models and datasets, for both positive and negative audio samples. The Universal threshold for each model is indicated. Weakly-supervised models are indicated in blue.

Modality Alignment and Separability. A good VSSL model should align auditory and visual features corresponding to a positive audio-visual pair and separate them in case of a negative pair, giving rise to a high value of audio-visual similarity in the first case and a low similarity in the second. We adopt the visualization of Figure 1 from [27], showing the distribution of the maximum values of the similarity maps of the positive and the different negative cases. As can be seen, SSL-SaN is the model that better learns to separate the positives from the negatives in general, considering the three datasets (the rest of models are

shown in the Supplementary material). We propose a new metric, denoted as *separability*:

$$\mathcal{S}ep = Q_1^+ - Q_3^- \tag{2}$$

where Q_1^+ denotes the 1st quartile of the maximum audio-visual similarities of the positive pairs and Q_3^- the 3rd quartile of the maximum audio-visual similarities of the negative pairs. $\mathcal{S}ep$ is a real number and the larger the value, the better the VSSL model is at discriminating positive audio-visual pairs from negative ones. A negative value indicates that the interquartile ranges of audio-visual similarity values for positive pairs ($Q_3^+ - Q_1^+$) and negative pairs ($Q_3^- - Q_1^-$) overlap. This last scenario is not desirable.

Test set	Model	Self Supervised	Magnitude ↓	Alignment ↑	Separability$_{\pm}$ ↑
VGG-SS	LVS [CVPR, 2021]	✘	*0.92*	0.19	-0.1124
	EZ-VSL [ECCV, 2022]	✘	1.14	0.20	-0.0339
	FNAC [CVPR, 2023]	✘	1.25	0.03	-0.0094
	SLAVC [NEURIPS, 2022]	✘	1.14	0.22	-0.0427
	SSL-Align [ICCV, 2023]	✘	0.94	*0.41*	*0.1034*
	ACL [WACV, 2025]	✘	1.31	0.14	-0.2551
	SSL-TIE [ACMMM, 2022]	✓	0.96	0.37	0.0606
	SSL-Align [ICCV, 2023]	✓	0.95	0.39	0.0428
	Ours → SSL-SaN	✓	**0.93**	**0.40**	**0.0971**
IS3	LVS [CVPR, 2021]	✘	*0.95*	0.11	-0.1113
	EZ-VSL [ECCV, 2022]	✘	1.14	0.17	-0.0423
	FNAC [CVPR, 2023]	✘	1.24	0.02	-0.0241
	SLAVC [NEURIPS, 2022]	✘	1.17	0.15	-0.0733
	SSL-Align [ICCV, 2023]	✘	0.98	*0.32*	-0.0608
	ACL [WACV, 2025]	✘	1.29	0.17	*0.1112*
	SSL-TIE [ACMMM, 2022]	✓	0.98	0.28	-0.0530
	SSL-Align [ICCV, 2023]	✓	0.99	0.26	-0.1287
	Ours → SSL-SaN	✓	**0.97**	**0.31**	**-0.0327**
IS3+	LVS [CVPR, 2021]	✘	*0.96*	0.10	-0.1148
	EZ-VSL [ECCV, 2022]	✘	1.14	0.17	-0.0423
	FNAC [CVPR, 2023]	✘	1.24	0.01	-0.0231
	SLAVC [NEURIPS, 2022]	✘	1.16	0.15	-0.0692
	SSL-Align [ICCV, 2023]	✘	0.98	*0.33*	-0.0480
	ACL [WACV, 2025]	✘	1.29	0.17	*0.0717*
	SSL-TIE [ACMMM, 2022]	✓	0.99	0.28	-0.0604
	SSL-Align [ICCV, 2023]	✓	0.99	0.26	-0.1279
	Ours → SSL-SaN	✓	**0.97**	**0.31**	**-0.0327**
AVS-Bench S4	LVS [CVPR, 2021]	✘	0.88	0.28	-0.1085
	EZ-VSL [ECCV, 2022]	✘	1.12	0.18	-0.0180
	FNAC [CVPR, 2023]	✘	1.23	0.01	-0.0017
	SLAVC [NEURIPS, 2022]	✘	1.15	0.14	-0.0327
	SSL-Align [ICCV, 2023]	✘	*0.86*	*0.49*	0.2161
	ACL [WACV, 2025]	✘	1.28	0.18	0.0575
	SSL-TIE [ACMMM, 2022]	✓	0.93	0.41	0.1442
	SSL-Align [ICCV, 2023]	✓	0.88	**0.48**	0.1729
	Ours → SSL-SaN	✓	***0.86***	0.47	***0.2479***

Table 1: Results of the cross-modal alignment and separability analysis. Best results in *italics*. Best results of self-supervised models in **bold** and the second-best underlined.

Table 1 reports the separability metric in VGG-SS, IS3, IS3+ and AVS-Bench S4 test sets. To better understand how models align both modalities in a positive pair, we also report the metrics of magnitude and alignment [17, 63]. *Alignment* measures how close the audio and image embeddings are for positive pairs using cosine similarity, while *magnitude* quantifies the L_2 distance between both embeddings. As shown in Table 1, the LVS model achieves the best scores in both magnitude and alignment in VGG-SS, IS3 and IS3$^+$, and second best in S4, which at first glance could suggest an excellent reduction of the modality gap. However, this result seems to indicate that the audio and image embeddings are collapsing into the same region of the embedding space, rather than aligning them in a semantically meaningful way. This collapse reduces their discriminative power and is reflected in LVS's poor performance on the separability metric. A reversed pattern is observed with FNAC, which achieves a relatively good separability score, but its poor alignment and high magnitude

suggest that the model struggles to structure the modality representations meaningfully. In contrast, SSL-SaN model achieves a better balance, obtaining the best results in magnitude and alignment among the self-supervised models in VGG-SS, IS3 and IS3$^+$, and best magnitude and second best alignment in S4. Most importantly, it achieves the best score, among the self-supervised models, in the separability metric in all test sets, which highlights a strong distinction between positive and negative samples. These results demonstrate the strength of our fully self-supervised training approach, which encourages the learning of semantically rich and well-structured audio-visual representations.

			Positive audio input				Negative audio input						Global metric	
			cIoU		AUC		Silence		Noise		Offscreen sound			
Test set	Model	Self Sup.	Uth ↑	Adap. ↑	Uth ↑	Adap. ↑	pIA ↓	AUC_N ↑	pIA ↓	AUC_N ↑	pIA ↓	AUC_N ↑	F_{LOC} ↑	F_{AUC} ↑
VGG-SS	RCGrad [INTERSPEECH, 2022]	✘	11.71	37.04	12.08	37.26	4.42	95.80	4.42	95.80	4.41	95.82	20.86	21.45
	LVS [CVPR, 2021]	✘	3.90	39.43	5.91	41.27	1.59	98.30	0.05	99.93	0.70	99.23	7.51	11.15
	EZ-VSL [ECCV, 2022]	✘	10.41	43.85	11.97	42.86	1.83	98.09	0.37	99.60	1.80	98.08	18.84	21.34
	FNAC [CVPR, 2023]	✘	17.45	47.14	18.69	44.27	4.09	95.88	4.09	95.88	3.58	96.38	29.53	31.28
	SLAVC [NEURIPS, 2022]	✘	9.60	49.62	11.40	45.55	5.53	94.43	0.59	99.39	1.54	98.45	17.48	20.40
	SSL-Align [ICCV, 2023]	✘	*34.46*	*56.86*	*34.78*	*49.25*	2.34	97.57	1.11	98.79	1.97	97.95	*51.02*	*51.36*
	ACL [WACV 2025]	✘	14.31	39.92	15.99	40.30	0.29	99.60	0.40	99.47	0.91	99.01	25.02	27.55
	SSL-TIE [ACMMM, 2022]	✓	27.78	51.88	28.23	48.00	0.78	99.16	0.68	99.26	2.50	97.39	43.36	43.90
	SSL-Align [ICCV, 2023]	✓	25.40	**53.92**	25.96	47.84	1.18	98.72	0.28	99.70	***0.63***	***99.33***	40.45	41.16
	Ours → SSL-SaN	✓	**29.61**	52.74	**29.97**	**48.66**	***0.01***	***99.99***	***0.00***	***100.00***	1.72	98.17	**45.63**	**46.05**
IS3	RCGrad [INTERSPEECH, 2022]	✘	0.10	1.07	2.54	3.57	4.49	95.81	4.49	95.81	4.47	95.79	0.20	4.95
	LVS [CVPR, 2021]	✘	2.07	14.10	4.31	26.37	1.08	98.80	0.03	99.95	*0.22*	*99.74*	4.05	8.27
	EZ-VSL [ECCV, 2022]	✘	3.82	16.82	5.88	28.53	1.10	98.80	0.01	99.98	0.78	99.14	7.35	11.11
	FNAC [CVPR, 2023]	✘	8.23	18.72	9.98	29.61	3.32	96.65	3.32	96.65	2.09	97.87	15.17	18.10
	SLAVC [NEURIPS, 2022]	✘	5.78	18.92	7.74	28.46	18.57	81.38	0.49	99.49	2.49	97.49	10.89	14.28
	SSL-Align [ICCV, 2023]	✘	25.25	34.61	25.96	39.23	4.33	95.49	1.17	98.67	1.95	97.94	40.12	41.00
	ACL [WACV 2025]	✘	*49.07*	*67.80*	*49.95*	*67.87*	0.02	99.86	1.37	98.78	0.50	99.40	*65.70*	*66.48*
	SSL-TIE [ACMMM, 2022]	✓	17.32	18.43	18.29	29.69	0.49	99.42	0.40	99.53	3.19	96.71	29.47	30.85
	SSL-Align [ICCV, 2023]	✓	17.37	**29.68**	18.54	**37.12**	3.33	96.45	0.55	99.42	**0.67**	**99.26**	29.54	31.20
	Ours → SSL-SaN	✓	**18.87**	18.90	**19.73**	30.28	***0.00***	***100.00***	***0.00***	***100.00***	1.97	97.87	**31.72**	**32.92**
IS3+	RCGrad [INTERSPEECH, 2022]	✘	0.10	1.07	2.54	3.57	4.50	52.49	4.49	52.49	4.48	52.49	0.20	4.95
	LVS [CVPR, 2021]	✘	2.05	14.35	4.31	26.72	1.08	98.80	0.03	99.95	*0.29*	*99.67*	4.01	8.26
	EZ-VSL [ECCV, 2022]	✘	3.91	17.95	5.95	28.70	1.16	98.75	0.02	99.98	0.95	98.95	7.52	11.23
	FNAC [CVPR, 2023]	✘	7.62	16.84	9.37	28.53	3.43	96.54	3.43	96.54	3.22	96.73	14.13	17.08
	SLAVC [NEURIPS, 2022]	✘	6.43	18.23	8.31	28.33	18.57	81.38	0.49	99.49	3.21	96.76	12.02	15.25
	SSL-Align [ICCV, 2023]	✘	25.63	33.36	26.29	38.11	4.33	95.49	1.17	98.67	2.30	97.58	40.58	41.39
	ACL [WACV 2025]	✘	*44.75*	*63.93*	*45.79*	*64.16*	0.02	99.86	1.37	98.79	0.39	99.64	*61.72*	*62.70*
	SSL-TIE [ACMMM, 2022]	✓	16.37	17.93	17.40	28.88	0.49	99.42	0.40	99.53	3.03	96.86	28.09	29.58
	SSL-Align [ICCV, 2023]	✓	17.11	**28.90**	18.30	**35.99**	3.33	96.45	0.55	99.42	**0.70**	**99.23**	29.15	30.86
	Ours → SSL-SaN	✓	**19.13**	18.90	**19.94**	30.28	***0.00***	***100.00***	***0.00***	***100.00***	2.09	97.73	**32.08**	**33.21**
AVS-Bench S4	RCGrad [INTERSPEECH, 2022]	✘	15.76	33.08	16.15	33.25	4.34	95.92	4.34	95.91	4.33	95.92	27.07	27.65
	LVS [CVPR, 2021]	✘	6.82	21.30	8.54	30.68	2.16	97.73	0.07	99.91	0.65	99.28	12.76	15.72
	EZ-VSL [ECCV, 2022]	✘	11.14	19.57	12.56	30.81	0.71	99.24	0.52	99.45	1.40	98.53	20.04	22.29
	FNAC [CVPR, 2023]	✘	18.57	22.85	19.55	33.14	5.79	94.17	5.79	94.17	3.17	96.81	31.07	32.42
	SLAVC [NEURIPS, 2022]	✘	9.07	22.58	10.80	33.13	3.72	96.24	1.17	98.81	1.42	98.55	16.60	19.45
	SSL-Align [ICCV, 2023]	✘	33.04	30.73	33.09	39.32	4.81	95.13	1.04	98.88	1.48	98.45	49.36	49.41
	ACL [WACV 2025]	✘	*51.34*	*65.75*	*49.57*	*64.11*	*0.00*	*99.99*	0.03	99.98	*0.32*	*99.72*	*67.82*	*66.26*
	SSL-TIE [ACMMM, 2022]	✓	28.40	31.24	28.64	38.60	2.90	97.08	2.69	97.25	1.37	98.56	44.00	44.29
	SSL-Align [ICCV, 2023]	✓	31.75	32.30	31.94	39.35	0.87	99.03	0.16	99.81	**0.48**	**99.49**	48.14	48.35
	Ours → SSL-SaN	✓	**32.76**	**33.87**	**32.86**	**41.11**	**0.05**	**99.95**	***0.00***	***100.00***	0.95	98.96	**49.31**	**49.42**

Table 2: Localization results on the VGG-SS, IS3$^+$ and S4 extended test sets. The best value across all models is shown in *italics*. Within the self-supervised subset, the best value is in **bold** and the second-best is underlined.

Localization results. As reported in Table 2, our model outperforms all prior self-supervised works in the three test sets in terms of cIoU-Uth, metrics of silence and noise, and more importantly, the global metrics. It is second best in offscreen metrics. For cIoU-Adap. it is the best in S4 and second best in the rest of datasets. Thanks to the addition of loss terms specifically addressing silence and noise during training, our model completely filters out silence and noise. Interestingly, the weakly-supervised version of SSL-Align achieves by far the best results on positive sounds, but not on the negative ones (across all datasets). In fact, it performs worse on negative sounds compared to its fully self-supervised version. We hypothesize that the weakly-supervised version, which uses a pretrained visual encoder from ImageNet, is more prone to leveraging objectness cues in the image [46]. This appears to be beneficial for positive sound but detrimental for negative ones. ACL achieves the best results in positive, offscreen and global metrics in IS3, IS3+ and S4. It leverages CLIPSeg [37], an

image segmentation model based on CLIP with a conditioned transformer decoder trained for object segmentation in a supervised way, with both positive and negative segmentation queries. Thus, it computes powerful and robust features for localization (highly related to segmentation) but less competitive for cross-modal retrieval, as shown in the next results.

Cross-modal retrieval. As in [65, 68], we evaluate VSSL models using a cross-modal retrieval task, to better assess how they capture the semantic correspondence between the audio and visual modalities. For this evaluation, we use Precision and Accuracy at top-K retrieved results, P@K and A@K, respectively. Precision refers to the proportion of the top-K retrieved items that belong to the same category as the query, while Accuracy considers a retrieval successful if at least one item among the top-K matches the query's category. Table 3 reports results for VGG-SS and AVS-Bench S4 (IS3 and IS3+ ones are in the Supp. mat.). Our model outperforms the self-supervised SSL-Align model at all K values in VGG-SS, IS3 and IS3$^+$ for both Precision and Accuracy in Image to Audio. Our model outperforms the self-supervised SSL-Align at high K for Precision in both Image to Audio and Audio to Image. On the other hand, the self-supervised version of SSL-Align outperforms our model in Audio to Image for all test sets except at high K values of Precision for S4 test set. Interestingly, our model improves by a large margin its baseline method, SSL-TIE, across all metrics and datasets, showing the benefits of including silence and noise in the learning stage.

Test set	Model	Self Supervised	I → A						A → I					
			P@1 ↑	P@5 ↑	P@10 ↑	A@1 ↑	A@5 ↑	A@10 ↑	P@1 ↑	P@5 ↑	P@10 ↑	A@1 ↑	A@5 ↑	A@10 ↑
VGG-SS	LVS [CVPR, 2021]	✘	2.96	2.84	2.83	2.96	10.37	16.47	4.02	3.92	3.54	4.02	13.71	20.22
	EZ-VSL [ECCV, 2022]	✘	2.11	2.27	2.16	2.11	8.21	13.06	3.94	3.51	3.24	3.94	14.19	22.55
	FNAC [CVPR, 2023]	✘	2.03	1.83	2.06	2.03	6.65	14.26	2.08	1.66	1.84	2.08	7.51	14.92
	SLAVC [NEURIPS, 2022]	✘	3.54	3.14	2.93	3.54	10.80	16.02	4.07	3.51	3.32	4.07	14.72	24.91
	SSL-Align [ICCV, 2023]	✘	*27.95*	*25.38*	*23.13*	*27.95*	*49.00*	58.64	*32.04*	*29.15*	*26.26*	*32.04*	*56.93*	*67.08*
	ACL [WACV, 2025]	✘	10.07	9.36	8.92	10.07	29.60	43.36	13.35	12.92	12.44	13.35	33.52	43.87
	SSL-TIE [ACMMM, 2022]	✓	16.25	14.87	13.83	16.25	35.81	47.51	15.12	14.82	13.81	15.12	35.91	47.41
	SSL-Align [ICCV, 2023]	✓	22.65	20.38	19.04	22.65	44.53	56.50	**26.67**	**23.82**	**21.40**	**26.67**	**51.26**	**61.85**
	Ours → SSL-SaN	✓	**24.61**	**22.94**	**20.88**	**24.61**	**48.57**	***58.97***	24.49	22.97	21.25	24.49	46.74	57.41
AVS-Bench S4	LVS [CVPR, 2021]	✘	20.81	23.89	25.15	20.81	56.49	70.81	33.38	30.81	30.26	33.38	54.59	66.22
	EZ-VSL [ECCV, 2022]	✘	2.97	3.70	3.81	2.97	13.65	23.92	5.14	5.14	4.86	5.14	5.41	9.19
	FNAC [CVPR, 2023]	✘	5.41	4.92	4.86	5.41	11.89	20.95	4.59	4.62	5.12	4.59	5.14	9.86
	SLAVC [NEURIPS, 2022]	✘	5.95	5.22	5.01	5.95	15.68	24.32	6.76	6.62	6.47	6.76	7.16	12.70
	SSL-Align [ICCV, 2023]	✘	*85.27*	*84.08*	*82.38*	*85.27*	*91.76*	93.24	*87.16*	*86.00*	*84.70*	*87.16*	*94.86*	*96.76*
	ACL [WACV, 2025]	✘	75.14	73.95	71.68	75.14	91.08	94.05	72.57	72.57	72.22	72.57	72.57	80.14
	SSL-TIE [ACMMM, 2022]	✓	73.38	75.27	74.92	73.38	85.27	90.00	66.76	66.76	66.69	66.76	66.76	77.03
	SSL-Align [ICCV, 2023]	✓	**80.00**	78.54	77.43	**80.00**	***91.76***	***94.59***	**81.89**	80.65	79.28	**81.89**	**91.08**	**94.73**
	Ours → SSL-SaN	✓	79.86	**80.81**	**81.50**	79.86	89.59	92.43	81.35	**81.38**	**81.82**	81.35	81.49	87.16

Table 3: Results of the cross-modal retrieval. Best results in *italics*. Best results of self-supervised models in **bold** and the second-best underlined.

Qualitative results. We show in Fig. 2 the localization maps of the self-supervised models for an example of IS3+. The similarity maps are filtered using the univeral threshold [27]. Unlike previous models, which fail to correctly localize both sound sources (dinosaur and firetruck) while filtering out silence, noise, and offscreen sounds, our approach successfully achieves both. More qualitative results (including failure cases and cross-modal retrieval results) are in the Supplementary material.

4.3 Ablation

First, we study the contribution of the new negative pairs with *silence* and *noise* as well as the loss terms $\mathcal{L}_S$ and $\mathcal{L}_N$. Table 4 shows this study on the AVS-Bench S4 test set. The best metrics, except for the case of offscreen sounds and separability, are obtained when considering silence, noise and $\mathcal{L}_S + \mathcal{L}_N$ during training. The best result for offscreen is achieved when considering only silence and $\mathcal{L}_S$, but this comes at the cost of worsening the results on positive sounds, silence and noise. The best result for separability is in the model

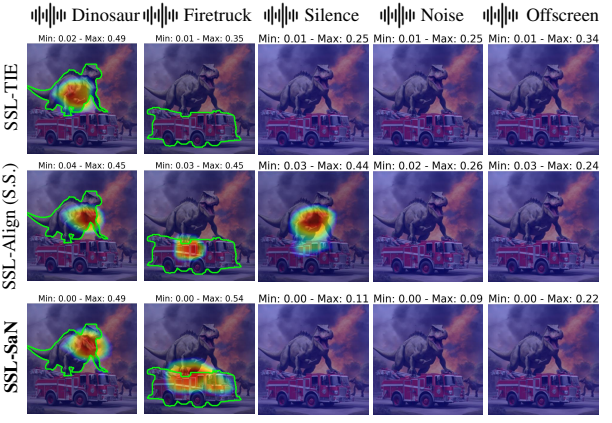

Figure 2: Localization results of different models in both *positive* and *negative* audio samples in the IS3+ test set. The audio-visual similarities below the universal threshold of each model are clipped, normalized to the interval $[0, 1]$, and overlaid with the original image. The boundaries of the ground-truth mask are shown in green. The minimum and maximum values of the audio-visual similarities are reported on top of each image.

trained with noise, but not with $\mathcal{L}_N$. The second best result for separability is obtained with the model using silence, noise and $\mathcal{L}_S + \mathcal{L}_N$. The results of this ablation on all test sets is present in the Supplementary material. Finally, Table A.6 in the Supplementary material shows an ablation study of the weight λ_{SN} that multiplies the new loss term $(\mathcal{L}_S + \mathcal{L}_N)$. Best results are achieved for $\lambda_{SN} = 1$.

Test set	S	N	$\mathcal{L}_S$	$\mathcal{L}_N$	cIoU$_{Uth}$ ↑	Negative audio input pIA$_S$ ↓	pIA$_N$ ↓	pIA$_O$ ↓	F$_{LOC}$ ↑	Sep ↑
AVS-Bench S4					31.57	2.53	2.49	1.00	47.76	0.2384
	✓				31.76	2.39	1.85	<u>0.82</u>	48.01	0.2350
	✓		✓		<u>32.49</u>	2.54	2.58	**0.81**	<u>48.80</u>	0.2458
		✓			31.97	2.26	2.59	0.88	48.22	**0.2558**
		✓		✓	32.07	2.21	2.26	1.07	48.34	0.2375
	✓	✓			32.05	<u>1.75</u>	<u>1.01</u>	0.84	48.40	0.2393
	✓	✓	✓	✓	**32.76**	**0.05**	**0.00**	0.95	**49.31**	<u>0.2479</u>

Table 4: Ablation table on the use of the silence (S) and noise (N) samples during training, and the new loss terms ($\mathcal{L}_S$ and $\mathcal{L}_N$). Best results in **bold**, second best <u>underlined</u>.

5 Conclusion

We present SSL-SaN, a simple yet effective approach that leverages silence and noise audio samples to enhance cross-modal retrieval, sound localization and robustness to negative audio samples. Additionally, we present IS3+, an improved version of IS3 that corrects mismatched image-audio pairs, improving evaluation reliability. Our model is comprehensively evaluated on four benchmark datasets using positive, negative, and global metrics. We further propose a new metric to quantify cross-modal alignment and feature separability, which also predicts sound localization and retrieval performance by capturing how well positive audio-visual pairs are distinguished from negative ones. Our approach outperforms recent state-of-the-art self-supervised models on most metrics and datasets.

Acknowledgements

The authors acknowledge support by Maria de Maeztu project ref. CEX2021-001195-M/AEI /10.13039/501100011033. X. J. has received financial support through FPI scholarship PRE2022-101321. G. H. has received financial support through Fulbright Program and Ministerio de Universidades (Spain) funding for mobility stays of professors and researchers in foreign higher education and research centers. M. F. has received support from the National Science Foundation under Grant Number 2504643. This work was supported in part through the NYU IT High Performance Computing resources, services, and staff expertise.

References

[1] Adobe. Adobe audition sound effects, 2023. URL https://www.adobe.com/products/audition/offers/adobeauditiondlcsfx.html. Accessed: [25-Jan-2025].

[2] 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*, pages 208–224. Springer, 2020.

[3] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *IEEE International Conference on Computer Vision*, pages 609–617, 2017.

[4] Relja Arandjelovic and Andrew Zisserman. Objects that sound. In *European Conference on Computer Vision*, pages 435–451, 2018.

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

[6] Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Localizing visual sounds the hard way. In *IEEE/CVF conference on computer vision and pattern recognition*, pages 16867–16876, 2021.

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607. PMLR, 2020.

[8] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021.

[9] Luyao Cheng, Hui Wang, Siqi Zheng, Yafeng Chen, Rongjie Huang, Qinglin Zhang, Qian Chen, and Xihao Li. Integrating audio, visual, and semantic information for enhanced multimodal speaker diarization. *arXiv preprint arXiv:2408.12102*, 2024.

[10] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, volume 1, pages 539–546, 2005.

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[12] Yuexi Du, Ziyang Chen, Justin Salamon, Bryan Russell, and Andrew Owens. Conditional generation of audio from video via foley analogies. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2426–2436, 2023.

[13] Amit Eliav and Sharon Gannot. Audio-visual approach for multimodal concurrent speaker detection. *arXiv preprint arXiv:2407.01774*, 2024.

[14] Stefan Fichna, Thomas Biberger, Bernhard U Seeber, and Stephan D Ewert. Effect of acoustic scene complexity and visual scene representation on auditory perception in virtual audio-visual environments. In *2021 Immersive and 3D Audio: from Architecture to Automotive (I3DA)*, pages 1–9. IEEE, 2021.

[15] John W Fisher III, Trevor Darrell, William Freeman, and Paul Viola. Learning joint statistical models for audio-visual fusion and segregation. *Advances in neural information processing systems*, 13, 2000.

[16] Ruohan Gao and Kristen Grauman. Visualvoice: Audio-visual speech separation with cross-modal consistency. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15490–15500. IEEE, 2021.

[17] Shashank Goel, Hritik Bansal, Sumit Bhatia, Ryan Rossi, Vishwa Vinay, and Aditya Grover. Cyclip: Cyclic contrastive language-image pretraining. *Advances in Neural Information Processing Systems*, 35:6704–6719, 2022.

[18] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

[19] Mark Hamilton, Andrew Zisserman, John R Hershey, and William T Freeman. Separating the" chirp" from the" chat": Self-supervised visual grounding of sound and language. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13117–13127, 2024.

[20] David R Hardoon, Sandor Szedmak, and John Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16 (12):2639–2664, 2004.

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

[22] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[23] John Hershey and Javier Movellan. Audio vision: Using audio-visual synchrony to locate sounds. *Advances in neural information processing systems*, 12, 1999.

[24] 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.

[25] Xixi Hu, Ziyang Chen, and Andrew Owens. Mix and localize: Localizing sound sources in mixtures. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10483–10492, 2022.

[26] Amir Jamaludin, Joon Son Chung, and Andrew Zisserman. You said that?: Synthesising talking faces from audio. *International Journal of Computer Vision*, 127:1767–1779, 2019.

[27] Xavier Juanola, Gloria Haro, and Magdalena Fuentes. A critical assessment of visual sound source localization models including negative audio. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–5, 2025.

[28] Einat Kidron, Yoav Y Schechner, and Michael Elad. Pixels that sound. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, volume 1, pages 88–95, 2005.

[29] Dongjin Kim, Sung Jin Um, Sangmin Lee, and Jung Uk Kim. Learning to visually localize sound sources from mixtures without prior source knowledge. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26467–26476, 2024.

[30] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[31] Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. *Advances in Neural Information Processing Systems*, 31, 2018.

[32] Yinjie Lei, Zixuan Wang, Feng Chen, Guoqing Wang, Peng Wang, and Yang Yang. Recent advances in multi-modal 3d scene understanding: A comprehensive survey and evaluation. *arXiv preprint arXiv:2310.15676*, 2023.

[33] Jia Li, Wenjie Zhao, Ziru Huang, Yunhui Guo, and Yapeng Tian. Do audio-visual segmentation models truly segment sounding objects? *arXiv preprint arXiv:2502.00358*, 2025.

[34] Susan Liang, Chao Huang, Yapeng Tian, Anurag Kumar, and Chenliang Xu. Av-nerf: Learning neural fields for real-world audio-visual scene synthesis. *Advances in Neural Information Processing Systems*, 36:37472–37490, 2023.

[35] Jinxiang Liu, Chen Ju, Weidi Xie, and Ya Zhang. Exploiting transformation invariance and equivariance for self-supervised sound localisation. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3742–3753, 2022.

[36] Xian Liu, Rui Qian, Hang Zhou, Di Hu, Weiyao Lin, Ziwei Liu, Bolei Zhou, and Xiaowei Zhou. Visual sound localization in the wild by cross-modal interference erasing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1801–1809, 2022.

[37] Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *IEEE/CVF conference on computer vision and pattern recognition*, pages 7086–7096, 2022.

[38] Tanvir Mahmud, Yapeng Tian, and Diana Marculescu. T-vsl: Text-guided visual sound source localization in mixtures. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26742–26751, 2024.

[39] Shentong Mo and Pedro Morgado. Localizing visual sounds the easy way. In *European Conference on Computer Vision*, pages 218–234. Springer, 2022.

[40] Shentong Mo and Pedro Morgado. A closer look at weakly-supervised audio-visual source localization. *Advances in Neural Information Processing Systems*, 35:37524–37536, 2022.

[41] Shentong Mo and Yapeng Tian. Audio-visual grouping network for sound localization from mixtures. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10565–10574, 2023.

[42] Juan F Montesinos, Venkatesh S Kadandale, and Gloria Haro. Vovit: Low latency graph-based audio-visual voice separation transformer. In *European Conference on Computer Vision*, pages 310–326. Springer, 2022.

[43] Juan F Montesinos, Daniel Michelsanti, Gloria Haro, Zheng-Hua Tan, and Jesper Jensen. Speech inpainting: Context-based speech synthesis guided by video. In *Interspeech*, pages 4459–4463, 2023.

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

[45] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *European Conference on Computer Vision*, pages 631–648, 2018.

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

[47] Sooyoung Park, Arda Senocak, and Joon Son Chung. Marginnce: Robust sound localization with a negative margin. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–5, 2023.

[48] Sooyoung Park, Arda Senocak, and Joon Son Chung. Can clip help sound source localization? In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5711–5720, 2024.

[49] 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.

[50] Janani Ramaswamy and Sukhendu Das. See the sound, hear the pixels. In *IEEE/CVF winter conference on applications of computer vision*, pages 2970–2979, 2020.

[51] Michael Risoud, J-N Hanson, Fanny Gauvrit, Christian Renard, P-E Lemesre, N-X Bonne, and Christophe Vincent. Sound source localization. *European annals of otorhinolaryngology, head and neck diseases*, 135(4):259–264, 2018.

[52] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[53] José Salas-Cáceres, Javier Lorenzo-Navarro, David Freire-Obregón, and Modesto Castrillón-Santana. Multimodal emotion recognition based on a fusion of audiovisual information with temporal dynamics. *Multimedia Tools and Applications*, pages 1–17, 2024.

[54] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4358–4366, 2018.

[55] 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, 2019.

[56] 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*, pages 4863–4867, 2022.

[57] Arda Senocak, Hyeonggon Ryu, Junsik Kim, and In So Kweon. Less can be more: Sound source localization with a classification model. In *IEEE/CVF conference on computer vision and pattern recognition*, pages 3308–3317, 2022.

[58] Arda Senocak, Hyeonggon Ryu, Junsik Kim, Tae-Hyun Oh, Hanspeter Pfister, and Joon Son Chung. Aligning sight and sound: Advanced sound source localization through audio-visual alignment. *arXiv preprint arXiv:2407.13676*, 2024.

[59] Zhaofeng Shi. A survey on audio synthesis and audio-visual multimodal processing. *arXiv preprint arXiv:2108.00443*, 2021.

[60] Zengjie Song, Jiangshe Zhang, Yuxi Wang, Junsong Fan, and Zhaoxiang Zhang. Enhancing sound source localization via false negative elimination. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[61] Weixuan Sun, Jiayi Zhang, Jianyuan Wang, Zheyuan Liu, Yiran Zhong, Tianpeng Feng, Yandong Guo, Yanhao Zhang, and Nick Barnes. Learning audio-visual source localization via false negative aware contrastive learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6420–6429, 2023.

[62] Kim Sung-Bin, Arda Senocak, Hyunwoo Ha, Andrew Owens, and Tae-Hyun Oh. Sound to visual scene generation by audio-to-visual latent alignment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6430–6440, 2023.

[63] Kim Sung-Bin, Arda Senocak, Hyunwoo Ha, and Tae-Hyun Oh. Sound2vision: Generating diverse visuals from audio through cross-modal latent alignment. *arXiv preprint arXiv:2412.06209*, 2024.

[64] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *European Conference on Computer Vision*, pages 247–263, 2018.

[65] Chenyun Wu, Zhe Lin, Scott Cohen, Trung Bui, and Subhransu Maji. Phrasecut: Language-based image segmentation in the wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10216–10225, 2020.

[66] Ho-Hsiang Wu, Magdalena Fuentes, Prem Seetharaman, and Juan Pablo Bello. How to listen? rethinking visual sound localization. *Interspeech*, 2022.

[67] Ehsan Yaghoubi, Andre Peter Kelm, Timo Gerkmann, and Simone Frintrop. Acoustic and visual knowledge distillation for contrastive audio-visual localization. In *Proceedings of the 25th International Conference on Multimodal Interaction*, pages 15–23, 2023.

[68] Yuhui Zhang, Jeff Z HaoChen, Shih-Cheng Huang, Kuan-Chieh Wang, James Zou, and Serena Yeung. Diagnosing and rectifying vision models using language. *arXiv preprint arXiv:2302.04269*, 2023.

[69] Jinxing Zhou, Xuyang Shen, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, and Yiran Zhong. Audio-visual segmentation with semantics. *International Journal of Computer Vision*, pages 1–21, 2024.