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

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.

Abstract

*Self-supervised audio-visual source localization aims to locate sound-source objects in video frames without extra annotations. Recent methods often approach this goal with the help of contrastive learning, which assumes only the audio and visual contents from the same video are positive samples for each other. However, this assumption would suffer from false negative samples in real-world training. For example, for an audio sample, treating the frames from the same audio class as negative samples may mislead the model and therefore harm the learned representations (e.g., the audio of a siren wailing may reasonably correspond to the ambulances in multiple images). Based on this observation, we propose a new learning strategy named False Negative Aware Contrastive (*FNAC*) to mitigate the problem of misleading the training with such false negative samples. Specifically, we utilize the intra-modal similarities to identify potentially similar samples and construct corresponding adjacency matrices to guide contrastive learning. Further, we propose to strengthen the role of true negative samples by explicitly leveraging the visual features of sound sources to facilitate the differentiation of authentic sounding source regions.* FNAC *achieves state-of-the-art performances on Flickr-SoundNet, VGG-Sound, and AVSBench, which demonstrates the effectiveness of our method in mitigating the false negative issue. The code is available at* https://github.com/OpenNLPLab/FNAC_AVL.

1. Introduction

When hearing a sound, humans can naturally imagine the visual appearance of the source objects and locate them in the scene. This demonstrates that audio-visual correspondence is an important ability for scene understanding. Given that unlimited paired audio-visual data exists in nature, there is an emerging interest in developing multi-modal systems with audio-visual understanding ability. Various audio-visual tasks have been studied, including sound source localization [8, 19–21, 26–28], audio-visual event localization [32, 33, 35, 39], audio-visual video parsing [11, 18, 31] and audio-visual segmentation [37, 38]. In this work, we focus on unsupervised visual sound source localization, with the aim of localizing the sound-source objects in an image using its paired audio clip, but without relying on any manual annotations.

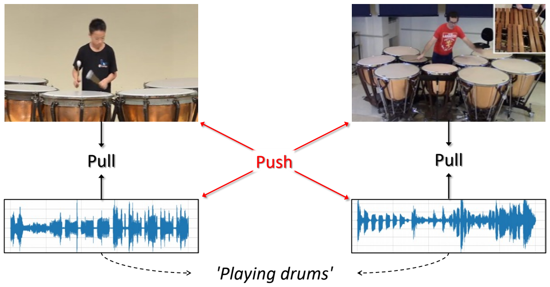

Figure 1. **False negative in audio-visual contrastive learning.** Audio-visual pairs with similar contents are falsely considered as negative samples to each other and pushed apart in the shared latent space, which we find would affect the model performance.

The essence of unsupervised visual sound source localization is to leverage the co-occurrences between an audio clip and its corresponding image to extract representations. A major part of existing methods [8, 19–21, 28] formulates this task as contrastive learning. For each image sample, its paired audio clip is viewed as the positive sample, while *all other* audio clips are considered as negative. Likewise, each audio clip considers its paired image as positive and *all others* as negative. As such, the Noise Contrastive Estimation (NCE) loss [24, 30] is used to perform instance discrimination by pushing closer the distance between a positive audio-image pair, while pulling away any negative pairs. However, the contrastive learning scheme above suffers from the issue of false negatives during training, *i.e.*, audio/image samples that belong to the semantically-matched class but are not regarded as a positive pair (due to the lack of manual labeling). A typical example is shown

[*] Indicates equal contribution

in Fig. 1. Research shows [4, 16, 29, 36] that these false negatives will lead to contradictory objectives and harm the representation learning.

Motivated by this observation, we assess the impact of false negatives in real-world training. We discover that with a batch size of 128, around 40% of the samples in VGG-Sound [9] will encounter at least one false negative sample during training. We then validate that false negatives indeed harm performance by artificially increasing the proportion of false negatives during training, and observing a noticeable performance drop. To make matters worse, larger batch sizes are often preferred in contrastive learning [24], but it may inadvertently increase the number of false negative samples during training and affect representation quality.

To this end, we propose a false-negative aware audio-visual contrastive learning framework (FNAC), where we employ the intra-modal similarities as weak supervision. Specifically, we compute pair-wise similarities between all audio clips in a mini-batch without considering the visual to form an audio intra-modal adjacency matrix. Likewise, in the visual modality, we obtain an image adjacency matrix. We found that the adjacency matrices effectively identify potential samples of the same class within each modality (Fig. 4). The information can then be used to mitigate the false negatives and enhance the effect of true pairings.

Specifically, we propose two complementary strategies: 1) **FNS** for False Negatives Suppression, and 2) **TNE** for True Negatives Enhancement. First, when optimizing the NCE loss, FNS regularizes the inter-modal and intra-modal similarities. Intrinsically, intra-modal adjacency explores potential false negatives by the similarity intensities and the pulling forces applied to these false negatives are canceled accordingly. Furthermore, we introduce TNE to emphasize the true negative influences in a region-wise manner, which in turn reduces the effect of false negative samples as well. We adopt the audio adjacency matrix to identify dissimilar samples, i.e., true negatives. Intuitively, dissimilar (true negative) sounds correspond to distinct regions, so the localized regions across the identified true negatives are regularized to be different. Such a mechanism encourages the model to discriminate genuine sound-source regions and suppress the co-occurring quiet objects. we conduct extensive analysis to demonstrate the effectiveness of our proposed method and report competitive performances across different settings and datasets. In summary, our main contributions are:

- We investigate the false negative issue in audio-visual contrastive learning. We quantitatively validate that this issue occurs and harms the representation quality.
- We exploit intra-modal similarities to identify potential false negatives and introduce FNS to suppress their impact.
- We propose TNE, which emphasizes true negatives using different localization results between the identified true negatives, thus encouraging more discriminative sound source localizations.

2. Related Work

False Negatives in Contrastive Learning. Typical contrastive learning employs instance discrimination [34] as a pretext task, in which two augmented views of the same image are considered as positive pairs, while views of all other images are treated as negative pairs, regardless of semantic similarities. Such a scheme inevitably suffers from the False Negatives issue [10, 13, 16, 36], which indicates that instances sharing the same semantic concepts are falsely treated as negatives, thus misleading model learning. Based on this, some works attempt to incorporate similar instances into model training to eliminate the impact of false negatives. For example, Zheng *et al.* [36] model a nearest neighbor graph for each batch of instances and execute a KNN-based multi-crop strategy to detect false negatives. Similarly, Dwibedi *et al.* [13] sample nearest neighbors from the dataset and treat them as positives for contrastive learning. More recently, [16] and [10] study how to identify false negatives without class labels and explicitly remove detected false negatives by two strategies, elimination and attraction, to improve contrastive loss. Other works use clustering-based methods to encode semantic structures [5, 7, 17] and then perform contrastive learning on these semantically similar cluster centers. In this paper, we propose and explore a similar problem in self-supervised audio-visual learning.

Self-Supervised Sound Source Localization. Sound source localization aims to learn to locate sound-source regions in videos. Recent approaches extensively leverage contrastive learning based on audiovisual correspondence to address this issue. For example, [2, 3, 28] adopt a dual-stream architecture to extract unimodal features respectively and then calculate a contrastive loss to update the audiovisual network. The final localization map is usually obtained by calculating the cosine similarity between audio and visual features. Following the paradigm, Mo *et al.* [21] further propose an object-guided localization (OGL) module, an extra pre-trained visual encoder, to introduce visual priors into the localization results. Nevertheless, these works assume that the paired audio-visual signals are regularly aligned and all mismatched samples are heterogeneous. As aforementioned, the assumption ignores the semantic similarities between samples. Accordingly, Chen *et al.* [8] incorporate explicitly background regions with low correlation to the given audio into the framework and regard them as *hard negatives*. In a slightly different task

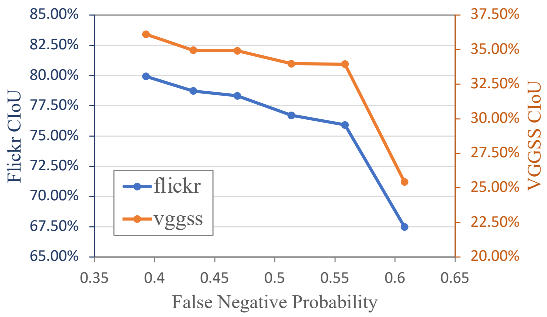

Figure 2. **Impact of false negatives on audio-visual representation learning**. We adopt Consensus Intersection over Union (CIoU) [28] as an evaluation metric (higher is better) and report results on FlickrSoundNet [6] and VGG-SS [9] test sets, depicted by blue and brown, respectively. An obvious performance decline is observed as the proportion of false negative samples increases.

setting of audio-visual instance discrimination, false negative issues are also investigated. [23] detects false negatives by defining sets of positive and negative samples via cross-modal agreement, where positive samples are defined as similar samples in both modalities. [22] considers both false positive and false negative issues. For false negatives, they estimate the similarity across instances to compute a soft target distribution over instances so the contributions of false negatives are down-weighted. One work that explores a similar problem to us is [29], which treats top-K semantically similar audio-visual pairs as *hard positives* and explicitly integrates them into the contrastive loss. Unfortunately, this method relies on manually selected hypermeters K and a hard threshold. Unlike the previous methods, this paper uses intra-modal adjacency matrices to adaptively detect *false negatives* and eliminate their impact.

3. Method

In this section, we first investigate the problem raised by false negative samples in Sec. 3.1, and then propose to mitigate the problem. From two aspects, without supervision we identify the false negative samples to explicitly suppress them in training (Sec. 3.2), and use region-wise comparing learning to enhance the roles of true negative samples which hence relatively suppresses false negatives (Sec. 3.2).

3.1. Revisiting Audio-visual Contrasting Learning

Various methods for audio-visual localization [18,21,28] employ a contrastive learning framework, in a manner of instance discrimination. Arguably, it assumes the audio-visual scenes from different videos are distinctive. Therefore the paired audio and image from the same video are considered a positive pair while the samples (audio or image) from other pairs are regarded as negatives. Specifically, we denote an audio-visual dataset as $\mathcal{D} \in \{(a_i, v_i), i \in [0, n)\}$, where (a_i, v_i) represents a sample with paired audio-visual content. Usually, a two-stream network is used to encode audio and visual signals and then map them into a shared latent space. The audio and visual representations extracted from (a_i, v_i) are denoted as $Z_i^a \in \mathbb{R}^{1\times d}$ and $Z_i^v \in \mathbb{R}^{1\times d}$ with d as feature dimension. Ideally, Z_i^a and Z_i^v represent the same semantic concept from the visual and audio perspectives, *e.g.*, playing the drum. The optimization objective of contrastive learning is to maximize the similarity between audio and visual representations from the same video while minimizing the similarity between features from different videos. Mathematically, it can be formulated in a modal-symmetric way for a pair (a_i, v_i) as:

$$\mathcal{L}_{\text{contrast\_i}} = -\log \frac{\exp\left[\frac{1}{\tau}\text{sim}(Z_i^a, Z_i^v)\right]}{\sum_j^b \exp\left[\frac{1}{\tau}\text{sim}(Z_i^a, Z_j^v)\right]} - \log \frac{\exp\left[\frac{1}{\tau}\text{sim}(Z_i^v, Z_i^a)\right]}{\sum_j^b \exp\left[\frac{1}{\tau}\text{sim}(Z_i^v, Z_j^a)\right]}, \quad (1)$$

where τ is a temperature hyper-parameter, b denotes batch size and sim represents the similarity function. Intuitively, this loss implies that each audio feature Z_i^a is pushed close to its paired visual feature Z_i^v in the shared latent space, while pulled apart from the rest $(b-1)$ visual features $Z_{j,j\neq i}^v$. However, as discussed, there exist $\hat{Z}_j^a$ and/or $\hat{Z}_j^v$ that are semantically similar to (Z_i^a, Z_i^v), *i.e.*, false negatives. In this situation, forcing $\text{sim}(Z_i^a, \hat{Z}_j^v)$ or $\text{sim}(Z_i^v, \hat{Z}_j^a)$ to be small might perplex the model training and lead to a non-optimal representation.

Impact of False Negatives. Based on this observation, we conduct several pilot experiments to verify the issue of false negatives and their influence on audio-visual representation learning. For simplification, we reasonably assume that samples that share the same manually labeled category are False Negatives. Firstly, we examine the data distribution during the training procedure. We find that over $\mathbf{39.27}\%$ samples suffer from at least one false negative sample when training with a batch size of 128 on the VGG-Sound dataset [9] covering 309 categories. This ratio will undoubtedly increase when employing bigger batch sizes or fewer categories. Secondly, we examine how the false negative issue might affect audio-visual localization performance. Following prior works [8, 21], we adopt ResNet-18 as the backbone to encode audio and visual features and a standard NCE loss [24] is used. In particular, we manually substitute the true negatives with false negatives in the training samples. As shown in Fig 2, when the false negative rate is progressively increased, a significant performance decline is found, indicating that the model is perplexed by these similar samples. The experiments above demonstrate that false negatives substantially impact the model quality and cannot be disregarded.

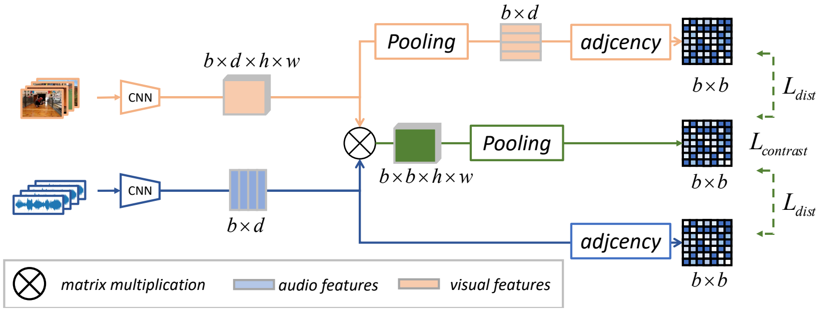

Figure 3. **An overview of False Negative Suppression (FNS).** The main optimization objective is NCE [24] loss on audio-visual pairs. The audio adjacency matrix and visual adjacency matrix are respectively constructed to suppress false negatives by regularizing NCE loss.

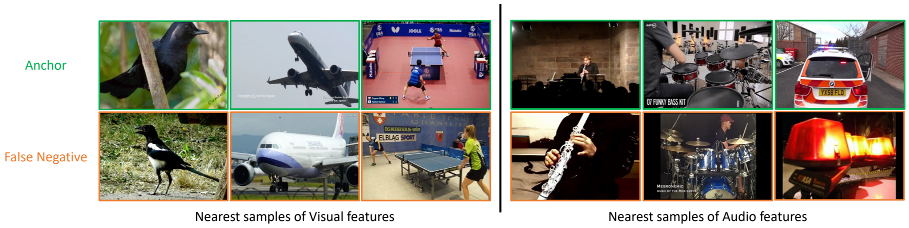
Anchor
False Negative
Nearest samples of Visual features
Nearest samples of Audio features

Figure 4. **Qualitative samples of the found potential false negatives in visual and audio modalities.** We report the nearest sample for every anchor. For the audio modality, we show the corresponding images here.

3.2. Mitigating False Negatives in Audio Visual Learning

As discussed in Sec. 3.1, we prove that the unpaired positive samples (*i.e.*, false negatives) existed in every mini-batch that may harm the representations. Correspondingly, we propose to solve this problem by FNAC with two complementary methods, False Negatives Suppression (FNS) and True Negatives Enhancement (TNE). For the same goal of mitigating the false negative issue, FNS identifies the false negative samples and regularizes to reduce their effects, while TNE enhances the contribution of true negatives by region-wise comparison, which also inadvertently suppresses false negatives. Both methods can be seamlessly integrated into the audio-visual contrastive framework as regularization terms.

FNS: False Negatives Suppression. To suppress the false negative effects, two challenges are posed. First, distinguishing the potential false negative samples within the current mini-batch without extra supervision such as class labels. Second, eliminating the misleading effects of the identified false negatives.

For the first challenge, we propose to leverage the unimodal feature representations to calculate pair-wise sample similarities, *i.e.*, adjacency matrix. As shown in Fig. 3, the audio clips are fed into an audio encoder to obtain audio features $Z^a \in \mathbb{R}^{b\times d}$, then we calculate a dot product with $(Z^a)^T$ followed by a row-wise softmax to obtain the pair-wise self-similarity matrix $S^a \in \mathbb{R}^{b\times b}$, *i.e.*, the audio adjacency matrix. Likewise, we average-pool the image features and obtain a visual adjacency matrix $S^v \in \mathbb{R}^{b\times b}$. Such adjacency matrices encode pair-wise similarities across a batch without considering inter-modal connection. For each audio-visual sample, we show its nearest sample by querying the adjacency matrices, as shown in Fig 4. It indicates that intra-modal adjacency matrices can effectively identify potential class-matched samples.

Regarding the second challenge, we propose to impose intra-modal adjacency as a soft supervision signal for inter-modal contrastive learning, *i.e.*, audio-visual similarities should be statistically consistent with the intra-modal similarities. In other words, if two scenes are close, their audio features similarity value $\text{sim}(Z_i^a, Z_j^a)$ should be high, the visual similarity $\text{sim}(Z_i^v, Z_j^v)$ should match, and the cross-modal similarities $\text{sim}(Z_i^a, Z_j^v)$ or $\text{sim}(Z_i^v, Z_j^a)$ should also show consistency. Therefore, we leverage intra-modal similarity as a supervision signal for audio-visual contrastive learning. Technically, we could utilize hard thresholds to achieve pseudo labels from intra-modal similarity like [8, 29, 36], which can serve to identify potential false negatives samples or areas. However, such methods require parameter tuning, and misassigned labels can exacerbate the false-negative problem. Differently, we propose FNS that directly regularizes the similarity scores. For a

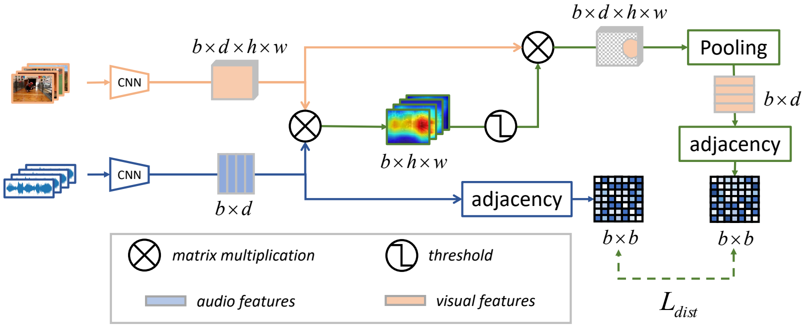

Figure 5. **An overview of True Negative Enhancement (TNE).** The visual features from localized regions are extracted to construct a sound-source visual feature adjacency matrix. The audio adjacency matrix is used to regularize the sound-source visual feature adjacency matrix to enhance true negatives.

sample i, its optimization objective is formulated as:

$$\mathcal{L}_{\text{FNS\_1}} = \frac{1}{b}\sum_{j}^{b} \mathcal{L}_{dist}(\text{sim}(Z_i^a, Z_j^v), \text{sim}(Z_i^a, Z_j^a)) \tag{2}$$

$$\mathcal{L}_{\text{FNS\_2}} = \frac{1}{b}\sum_{j}^{b} \mathcal{L}_{dist}(\text{sim}(Z_i^a, Z_j^v), \text{sim}(Z_i^v, Z_j^v)) \tag{3}$$

For an audio sample a_i, the NCE loss $L_{\text{contrast\_i}}$ in Eq. 1 pushes all negatives away by encouraging $\text{sim}(Z_i^a, Z_j^v)_{j\in[0,n),j\neq i}$ to be close to 0, whereas FNS pulls back false negatives to suppress their effects. The suppression intensity corresponds to the values of $\text{sim}(Z_i^a, Z_j^a)$ and $\text{sim}(Z_i^v, Z_j^v)$. In Eq. 3, the visual contrastive loss is calculated symmetrically. In practice, we calculate L1 distances between the inter-modal contrastive matrix and intra-modal adjacency matrices as shown in Fig. 3, so all pair-wise similarities are considered. Such a regularization term yields a parameter-free process so that we do not need to choose a hard threshold to determine the real false negatives, as the false negative effects can be adaptively regularized.

TNE: True Negatives Enhancement. To further reduce the misleading effect of false negatives, we propose to enhance the contribution of true negatives by region-wise comparison. As opponents, the impact of true negatives and that of false negatives are relative, so if the role of true negatives is raised, the role of false negatives would be suppressed. To improve the effect of true negatives in audio-visual learning, we turn back to its core concept, *i.e.*, that the *sound source objects* are different both audibly and visually between the true negatives. Therefore, it is straightforward to put a particular emphasis on the possible regions of genuine sound-emitting objects.

Specifically, we localize the sound-source objects by paired audio-visual samples, pop up their regional visual features, and encourage those of true negative samples to be pulled away. The process is called true negative enhancement (TNE). It is worth noting that similar regularizations have been mentioned but disregarded by previous methods [8,21,29] because they did not distinguish the true negatives and false negatives samples.

We show the paradigm of TNE in Fig. 5. Given an audio-visual pair (a_i, v_i), we obtain its localization result and use the localization map as a mask to extract the localized visual representation $Z_i^s \in \mathbb{R}^{d\times h\times w}$, where s indicates it is sounding region visual representation. In other words, Z_i^s denotes the visual features that are aligned with the paired audio features. Then, consider another arbitrary audio-visual pair (a_j, v_j) which localizes visual features Z_j^s. If a_i and a_j are semantically different, i.e., a_j is a true negative of a_i, then the Z_i^s should be dissimilar to Z_j^s. It encourages the model to focus on different pixels so as to mine discriminative visual features according to the audio similarities. To leverage such a constraint in practice, we regularize the audio adjacency matrix and the similarities between the sounding region visual features. Formally, the TNE regularization is:

$$\mathcal{L}_{\text{TNE}} = \frac{1}{b}\sum_{j}^{b} \mathcal{L}_{dist}(\text{sim}(Z_i^a, Z_j^a), \text{sim}(Z_i^s, Z_j^s)) \tag{4}$$

FNS and TNE are two general mechanisms for multi-modal contrastive learning. FNS adopts audio and visual adjacency matrices to explore the potential false negatives and suppress their misleading effects on the NCE loss. TNE uses audio adjacency to discriminate the sound-source localization so the model tends to discover the genuine sound sources. Both methods can be seamlessly integrated with the existing contrastive learning framework as extra regularization terms. Our final optimization objective for the i-th sample pair is:

$$\mathcal{L}_i = \mathcal{L}_{\text{contrast\_i}} + \alpha\mathcal{L}_{\text{FNS\_1}} + \beta\mathcal{L}_{\text{FNS\_2}} + \gamma\mathcal{L}_{\text{TNE}} \tag{5}$$

Table 1. Quantitative results of the model trained with Flickr 10k and 144k. Note that 'EZ-VSL + OGL' corresponds to the main results reported in [21]. 'EZ-VSL' indicates our reproduced results without OGL, which are not reported in [21]. We reproduce the results with the trained weights and code provided by [21].

Train set	Method	Flickr CIoU(%)	Flickr AUC(%)	VGG-SS CIoU(%)	VGG-SS AUC(%)
Flickr 10k	Attention10k [28]	43.60	44.90	-	-
	CoursetoFine [26]	52.20	49.60	-	-
	AVObject [1]	54.60	50.40	-	-
	LVS [8]	58.20	52.50	-	-
	EZ-VSL* [21]	62.24	54.74	19.86	30.96
	Ours	**84.33**	**63.26**	**35.27**	**38.00**
	EZ-VSL + OGL [21]	81.93	62.58	37.61	39.21
	Ours + OGL	**84.73**	**64.34**	**40.97**	**40.38**
Flickr 144k	Attention10k [28]	66.00	55.80	-	-
	DMC [15]	67.10	56.80	-	-
	LVS [8]	69.90	57.30	-	-
	HardPos [29]	75.20	59.70	-	-
	EZ-VSL* [21]	72.69	58.70	30.27	35.92
	Ours	**78.71**	**59.33**	**33.93**	**37.29**
	EZ-VSL + OGL [21]	83.13	63.06	41.01	40.23
	Ours + OGL	**83.93**	**63.06**	**41.10**	**40.44**

where α, β, γ are hyperparameters.

4. Experiments

4.1. Experimental Settings

Datasets We train our audio-visual localization model on two datasets: Flickr SoundNet [6] and VGG-Sound [9]. Flickr SoundNet contains 2 million unconstrained videos from Flickr. For a fair comparison with the existing methods [8, 20, 21, 29], we conduct the training on two subsets of 10k and 144k paired samples from Flicker SoundNet. VGG-Sound includes 200k video clips from 309 classes. We also train on two subsets of 10k and 144k paired samples following the convention.

Localization performances are measured on four benchmarks, Flickr [6], VGG-SS [9], Heard 110 and AVS-Bench [38]. The Flickr test set has 250 audio-visual pairs with manually labeled bounding boxes. VGG-SS is more challenging with 5,000 audio-visual pairs over 220 categories. Heard 110 is another subset of VGG-Sound to test the open-set learning ability. Its train set has 110 classes and the *val* set has another disjoint 110 unheard/unseen classes. Finally, AVSBench [38] is a recently proposed audio-visual dataset with 5,356 videos over 23 classes. It provides pixel-wise labels for fine-grained localization evaluation.

Implementation Details We implement our method with PyTorch. The images are resized and randomly cropped into 224×224 resolution, together with random horizontal flipping. The audio inputs are extracted from 3 seconds of audio clips and converted into log spectrogram maps. We also apply audio augmentation including Frequency mask and Time mask [25]. For both visual and audio encoders, we adopt ResNet18 [14] and the visual encoder is pre-trained on ImageNet-1k [12]. The model is optimized for 30 epochs with Adam using a learning rate of 10^{-4} and a weight decay of 10^{-4}. To achieve a stable representation, we warm up the network with only NCE loss for 3 epochs, then integrate our regularization for the remaining epochs.

4.2. Comparison on Flickr-SoundNet and VGG-SS

Flickr-SoundNet We train our model on Flickr 10k and 144k and report performances in Table 1. FNAC achieves superior performances over previous methods on both Flickr and VGG-SS test sets. Notably, compared to previous state-of-the-art EZ-VSL [21] on the Flickr test set, FNAC achieves a striking improvement of 22.09% CIoU with 10k training samples and 6.02% CIoU with 144k training samples. When tested on the more challenging VGG-SS, FNAC also outperforms EZ-VSL by 15.41% CIoU and 3.49% CIoU respectively. Finally, Object-Guided localization (OGL) is a post-processing strategy that adopts pure visual-based localization results to refine audio-visual localization. For a fair comparison, we integrate FNAC with OGL to compare with EZ-VSL which also uses OGL. As shown, FNAC also outperforms EZ-VSL in most cases.

VGG-SS Performance of the model trained on VGG-Sound 10k and 144k is reported in Table 2. FNAC also beats all previous methods with a clear margin. For example, on the VGG-SS test set, we outperform EZ-VSL by 11.45% CIoU with 10k and 5.31% CIoU with 144k.

We highlight two results as demonstrated in Table 1 and Table 2. First, FNAC achieves similar results on 10k and 144k training sets in both Flickr SoundNet and VGGSound, which is not observed in previous models. We believe that small-scale datasets (Flickr 10k, VGG-Sound 10k) have fewer semantic classes and will encounter more false negatives during training, so previous methods have substantial performance gaps between the 10k and 144k training

Table 2. Quantitative results of models trained with VGG-SS 10k and 144k.

Train set	Method	Flickr CIoU(%)	Flickr AUC(%)	VGG-SS CIoU(%)	VGG-SS AUC(%)
VGGSound 10k	LVS [8]	61.80	53.60	-	-
	EZ-VSL* [21]	63.85	54.44	25.84	33.68
	Ours	**85.74**	**63.66**	**37.29**	**38.99**
	EZ-VSL + OGL [21]	78.71	61.53	38.71	39.80
	Ours + OGL	**82.13**	**63.64**	**40.69**	**40.42**
VGGSound 144k	Attention10k [28]	-	-	18.50	30.20
	DMC [15]	-	-	29.10	34.80
	AVObject [1]	-	-	29.70	35.70
	LVS [8]	73.50	59.00	34.40	38.20
	HardPos [29]	76.80	59.20	34.60	38.00
	EZ-VSL [21]	79.51	61.17	34.38	37.70
	Ours	**84.73**	**63.76**	**39.50**	**39.66**
	EZ-VSL + OGL [21]	83.94	63.60	38.85	39.54
	Ours + OGL	**85.14**	**64.30**	**41.85**	**40.80**

Table 3. Quantitative results on Heard 110 and Unheard 110. For a fair comparison, the results of EZ-VSL [21] and ours are integrated with the OGL module.

Test Set	Method	CIoU(%)	AUC(%)
Heard 110	LVS [8]	28.90	36.20
	EZ-VSL [21]	37.25	38.97
	Ours	**39.54**	**39.83**
Unheard 110	LVS [8]	26.30	34.70
	EZ-VSL [21]	39.57	39.60
	Ours	**42.91**	**41.17**

Table 4. Zero-shot results on AVSBench S4 and MS3 [38]. All models are pretrained on VGGSound-144k dataset.

Test set	Method	mIoU	FScore
S4	LVS	23.69	.251
	EZ-VSL	26.43	.292
	Ours	**27.15**	**.314**
MS3	LVS	18.54	.174
	EZ-VSL	21.36	.216
	Ours	**21.98**	**.225**

Table 5. Analysis of each component of the proposed FNAC. aud adj: audio adjacency matrix. img adj: image adjacency matrix.

FNS		TNE	Flickr CIoU(%)	VGG-SS CIoU(%)
aud adj	img adj			
			77.91	33.93
✓			79.91	35.85
	✓		81.92	36.58
✓	✓		84.33	36.92
		✓	81.12	36.04
✓	✓	✓	**85.74**	**37.29**

sets. It indicates that FNAC can effectively address the false negative issue and has a strong ability for representation learning on small-scale datasets. Second, we show cross-dataset evaluation results in both tables, i.e., train on Flickr and test on VGG-SS or vice versa. FNAC achieves strong results and outperforms existing methods, which validates the cross-dataset generalization ability of FNAC. We show qualitative localization results in Fig. 6.

4.3. Comparison on Heard 110 and AVSBench

Heard 110 To assess the generalization ability of FNAC in unseen/unheard audiovisual scenes, we conduct an open set experiment. We use the 70k samples covering 110 categories randomly sampled from VGGSound for training and then evaluate the model on the same 110 heard categories and another disjoint set with 110 unheard categories. As shown in Table 3, FNAC considerably outperforms previous methods, especially on Unheard 110 (42.91% *vs.* 39.57% of CIoU), which demonstrates the generalization ability of FNAC in unconstrained audio-visual data.

AVSBench AVSBench [38] is a newly proposed audio-visual segmentation benchmark with pixel-level annotations, which can be regarded as a fine-grained sound source localization task and used to accurately evaluate the localization ability of models. We directly perform a zero-shot evaluation on the AVSbench with metrics of mIoU and F-Score in two settings, Single Sound Source Segmentation (S4) and Multiple Sound Source Segmentation (MS3), without any fine-tuning. The results are reported in Table 4. The proposed method achieves outstanding performance on both S4 and MS3 settings, *e.g.*, 27.15% and 21.98% mIoU when trained on VGGSound 144k. These results on the fine-grained localization benchmark validate the effectiveness of the TNE, which enables the model to discriminate authentic sound-source regions.

4.4. Ablation Analysis of FNAC

We propose different regularization terms to guide audio-visual contrastive learning. In Table 5, we ablate each component individually. All results are obtained following the same hyperparameter setting on VGG-SS 10k and the baseline is trained with only NCE loss. First, the audio adjacency matrix and visual adjacency matrix are obtained in FNS to suppress the false negatives. We show that each adjacency matrix can improve the performances over the

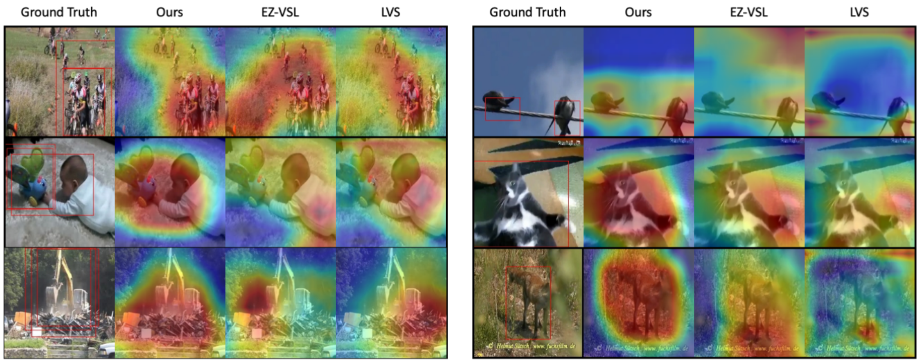

Figure 6. Visualization comparison on Flickr-SoundNet (left) and VGG-SS (right) test sets.

baseline, which indicates that potential false negatives can be effectively suppressed in both audio and visual modalities. By combining the two adjacency matrices together, FNS achieves 84.33% CIoU on Flickr and 36.92% CIoU on VGG-SS. Second, by only deploying TNE, we achieve improved results over baseline (81.12% CIoU on Flickr and 36.04% CIoU on VGG-SS), indicating the effectiveness of TNE in enhancing true negatives. Finally, the combination of FNS and TNE achieves significant improvement over the baseline, showing the effectiveness of FNAC. We refer readers to supp. mat. for further ablation studies.

4.5. Mining the Potential False Negatives

In this section, we show quantitative and schematic analysis of the false negative mining ability of our method. Intuitively, a good audio-visual model should be able to generate similar feature representations for class matched samples. To validate this intuition, we construct a batch where all samples belong to the same category, i.e., all false negatives, and show the audio-visual similarity matrices of EZ-VSL [21] and ours in Fig. 7. As shown in the left, EZ-VSL only highlights the diagonal, since each audio feature is only similar to the paired image feature while all others are dissimilar despite all samples being from the same class. Differently, the similarity distribution of our matrix is more evenly spread as the majority of the samples are regarded as similar. It indicates that FNAC has implicitly learned semantically-aware features and clustered them in the latent space, so false negatives will be effectively identified as they are closer in terms of feature distance.

Further, we show quantitative results of audio-visual similarities in Table 6 . When the batch contains all true negatives, our average similarity score is lower than previous methods [8, 21]. When the batch contains all false negatives, our average similarity score is higher. The margin between the two similarities demonstrates our ability to distinguish false negatives and true negatives.

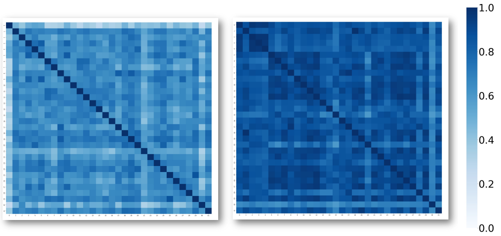

Figure 7. Cross-modal similarity matrix predicted by EZ-VSL (left) and ours (right) when all samples in the batch belong to the same category, *namely*, they are false negatives of each other. All values are normalized between 0 to 1.

Table 6. Audio-visual similarities with different data. TN: all samples in the batch belong to different categories. FN: all samples in the batch belong to the same category.

Method	TN ↓	FN ↑
LVS	0.4484	0.5102
EZ-VSl	0.5858	0.5938
Ours	**0.3812**	**0.6554**

5. Conclusion

In this paper, we propose a simple yet effective strategy named FNAC to deal with the false negative issue in audio-visual sound source localization. We propose two complementary strategies to suppress false negatives (FNS) and enhance true negatives (TNE). The intra-modal adjacency matrices of audio and visual are generated to identify false negatives. Then two regularization terms are seamlessly incorporated with contrastive learning, which enable the model to learn semantic-aware features from audio-visual pairs. We show that FNAC can effectively mitigate the false negative issues and achieve state-of-the-art performance on several audio-visual localization benchmarks. We hope that our method can facilitate future research in audio-visual learning and other multi-modal tasks.

References

[1] Triantafyllos Afouras, Andrew Owens, Joon Son Chung, and Andrew Zisserman. Self-supervised learning of audio-visual objects from video. In *European Conference on Computer Vision*, pages 208–224. Springer, 2020. 6, 7

[2] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 609–617, 2017. 2

[3] Relja Arandjelovic and Andrew Zisserman. Objects that sound. In *Proceedings of the European conference on computer vision (ECCV)*, pages 435–451, 2018. 2

[4] Sanjeev Arora, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. A theoretical analysis of contrastive unsupervised representation learning. *arXiv preprint arXiv:1902.09229*, 2019. 2

[5] YM Asano, C Rupprecht, and A Vedaldi. Self-labelling via simultaneous clustering and representation learning. arxiv 2019. *arXiv preprint arXiv:1911.05371*, 2019. 2

[6] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. *Advances in neural information processing systems*, 29, 2016. 3, 6

[7] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European conference on computer vision (ECCV)*, pages 132–149, 2018. 2

[8] 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. 1, 2, 3, 4, 5, 6, 7, 8

[9] 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. 2, 3, 6

[10] Tsai-Shien Chen, Wei-Chih Hung, Hung-Yu Tseng, Shao-Yi Chien, and Ming-Hsuan Yang. Incremental false negative detection for contrastive learning. *arXiv preprint arXiv:2106.03719*, 2021. 2

[11] Haoyue Cheng, Zhaoyang Liu, Hang Zhou, Chen Qian, Wayne Wu, and Limin Wang. Joint-modal label denoising for weakly-supervised audio-visual video parsing. *arXiv preprint arXiv:2204.11573*, 2022. 1

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

[13] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9588–9597, 2021. 2

[14] 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. 6

[15] 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. 6, 7

[16] Tri Huynh, Simon Kornblith, Matthew R Walter, Michael Maire, and Maryam Khademi. Boosting contrastive self-supervised learning with false negative cancellation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2785–2795, 2022. 2

[17] Junnan Li, Pan Zhou, Caiming Xiong, and Steven CH Hoi. Prototypical contrastive learning of unsupervised representations. *arXiv preprint arXiv:2005.04966*, 2020. 2

[18] Yan-Bo Lin, Hung-Yu Tseng, Hsin-Ying Lee, Yen-Yu Lin, and Ming-Hsuan Yang. Exploring cross-video and cross-modality signals for weakly-supervised audio-visual video parsing. *Advances in Neural Information Processing Systems*, 34:11449–11461, 2021. 1, 3

[19] 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

[20] 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. 1, 6

[21] Shentong Mo and Pedro Morgado. Localizing visual sounds the easy way. *arXiv preprint arXiv:2203.09324*, 2022. 1, 2, 3, 5, 6, 7, 8

[22] Pedro Morgado, Ishan Misra, and Nuno Vasconcelos. Robust audio-visual instance discrimination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12934–12945, 2021. 3

[23] Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. Audio-visual instance discrimination with cross-modal agreement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12475–12486, 2021. 3

[24] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 1, 2, 3, 4

[25] 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. 6

[26] 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. 1, 6

[27] Janani Ramaswamy and Sukhendu Das. See the sound, hear the pixels. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2970–2979, 2020. 1

[28] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *Proceedings of the IEEE Conference*

on Computer Vision and Pattern Recognition, pages 4358–4366, 2018. 1, 2, 3, 6, 7

[29] Arda Senocak, Hyeonggon Ryu, Junsik Kim, and In So Kweon. Learning sound localization better from semantically similar samples. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4863–4867. IEEE, 2022. 2, 3, 4, 5, 6, 7

[30] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *European conference on computer vision*, pages 776–794. Springer, 2020. 1

[31] Yapeng Tian, Dingzeyu Li, and Chenliang Xu. Unified multisensory perception: Weakly-supervised audio-visual video parsing. In *European Conference on Computer Vision*, pages 436–454. Springer, 2020. 1

[32] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 247–263, 2018. 1

[33] Yu Wu, Linchao Zhu, Yan Yan, and Yi Yang. Dual attention matching for audio-visual event localization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6292–6300, 2019. 1

[34] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018. 2

[35] Jiashuo Yu, Ying Cheng, Rui-Wei Zhao, Rui Feng, and Yuejie Zhang. Mm-pyramid: multimodal pyramid attentional network for audio-visual event localization and video parsing. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 6241–6249, 2022. 1

[36] Mingkai Zheng, Fei Wang, Shan You, Chen Qian, Changshui Zhang, Xiaogang Wang, and Chang Xu. Weakly supervised contrastive learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10042–10051, 2021. 2, 4

[37] Jinxing Zhou, Xuyang Shen, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, et al. Audio-visual segmentation with semantics. *arXiv preprint arXiv:2301.13190*, 2023. 1

[38] Jinxing Zhou, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, and Yiran Zhong. Audio-visual segmentation. In *European Conference on Computer Vision*, 2022. 1, 6, 7

[39] Jinxing Zhou, Liang Zheng, Yiran Zhong, Shijie Hao, and Meng Wang. Positive sample propagation along the audio-visual event line. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8436–8444, 2021. 1