Tianyu Liu Northwestern Polytechnical University Xi'an, Shanxi, China Ningbo Institute of Northwestern Polytechnical University Ningbo, Zhejiang, China reallty@mail.nwpu.edu.cn

Yufei Zha Northwestern Polytechnical University Xi'an, Shanxi, China Ningbo Institute of Northwestern Polytechnical University Ningbo, Zhejiang, China yufeizha@nwpu.edu.cn Peng Zhang* Northwestern Polytechnical University Xi'an, Shanxi, China Ningbo Institute of Northwestern Polytechnical University Ningbo, Zhejiang, China zh0036ng@nwpu.edu.cn

Tao You Northwestern Polytechnical University Xi'an, Shanxi, China youtao@nwpu.edu.cn Wei Huang Nanchang University Nanchang, Jiangxi, China huangwei@ncu.edu.cn

Yanning Zhang Northwestern Polytechnical University Xi'an, Shanxi, China ynzhang@nwpu.edu.cn

ABSTRACT

Self-supervised sound source localization is usually challenged by the modality inconsistency. In recent studies, contrastive learning based strategies have shown promising to establish such a consistent correspondence between audio and sound sources in visual scenarios. Unfortunately, the insufficient attention to the heterogeneity influence in the different modality features still limits this scheme to be further improved, which also becomes the motivation of our work. In this study, an Induction Network is proposed to bridge the modality gap more effectively. By decoupling the gradients of visual and audio modalities, the discriminative visual representations of sound sources can be learned with the designed Induction Vector in a bootstrap manner, which also enables the audio modality to be aligned with the visual modality consistently. In addition to a visual weighted contrastive loss, an adaptive threshold selection strategy is introduced to enhance the robustness of the Induction Network. Substantial experiments conducted on SoundNet-Flickr and VGG-Sound Source datasets have demonstrated a superior performance compared to other state-of-the-art works in different challenging scenarios. The code is available at https://github.com/Tahy1/AVIN.

CCS CONCEPTS

• Computing methodologies → Artificial intelligence.

*Corresponding author.

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada.

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM. ACM ISBN 979-8-4007-0108-5/23/10...\$15.00 https://doi.org/10.1145/3581783.3612502

KEYWORDS

audio-visual; sound source localization; contrastive learning; modality gap

ACM Reference Format:

Tianyu Liu, Peng Zhang, Wei Huang, Yufei Zha, Tao You, and Yanning Zhang. 2023. Induction Network: Audio-Visual Modality Gap-Bridging for Self-Supervised Sound Source Localization. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29–November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3581783.3612502

1 INTRODUCTION

Audio-visual Sound Source Localization (AV-SSL) can fundamentally support the intelligent Human-Computer Interaction (HCI) [8] by imitating the perceptive connection of human beings. To bridge the consistency between different modalities for sensing strengthen, the aggregation/combination of distinct modality representations has been employed [1, 12, 24] for representation alignment, but the intrinsic disparities in modalities usually limit such a capability of AV-SSL to achieve more robust performance in a variety of scenarios.

Comparatively, the metric learning has been adopted in recent studies to acquire uniform audio-visual representations for modal semantics synchronization, e.g. cosine similarity calculation of audio-visual modalities [18], which can benefit the localization of sound sources. By aligning the audio and visual modalities, some approaches choose to fuse the output features of a modality-specific encoder and minimize an objective function, such as Information Noise Contrastive Estimation (InfoNCE) loss or cross-entropy loss, for a proxy task [51]. A representative work based on modality alinement is proposed by Senocak et al. [34], which locate sound sources by learning the relationship between audio and global visual features. In a different way, Chen et al [5]. utilizes the pixel-wise

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Tianyu Liu, Peng Zhang, Wei Huang, Yufei Zha, Tao You, and Yanning Zhang



Figure 1: Two approaches of audio-visual representation alignment. (a) Current solutions align audio and visual modalities in feature space directly. (b) Our AVIN extract semantics from visual modality and align with audio modality in semantic space.

audio-visual correspondence to establish the connection between audio and local visual features. With a thresholding operation to categorize the features into positive, negative, and ignore regions, [5] proposes to maximize the cosine similarity of representations between the audio and the region of sound source in visual modality.

However, since the heterogeneity of different modalities has been widely disregarded in current solutions (align in feature space directly, as illustrated in Figure 1 (a)), the gap of audio-visual modalities still challenges the connection between the sounding objects with corresponding sound sources. Furthermore, the gradients coupling of distinct modalities also limits an enhancement of performance.

In this work, an Audio-Visual Induction Network (AVIN) is introduced to achieve more effective sound source localization. As illustrated in Figure 1 (b), the main operations in the proposed AVIN involves the semantic information extraction from the modality in a low-cost manner, as well as the representations alignment of different modalities based on the semantic information.

Assuming that the sounding objects exist in the image, the proposed Induction Network is capable of adequately exploiting the spatial information of visual modalities. Rather than enforcing the alignment of audio and visual modalities directly, the visual modality is inducted to distill the representation of the complete sounding object in a bootstrapped manner, which is then followed by the alignment of audio modality.

Furthermore, we have discovered that the stop-gradient (stopgrad) can significantly benefit the learning of consistent audiovisual representation. In more recent self-supervised pre-training tasks [7, 14], the stop-grad has been widely used to prevent the representation collapse of Siamese networks, but it is unlike the two-stream audio-visual networks that does not typically collapse into a constant. Nonetheless, the coupled gradient of audio-visual modalities during learning still makes the back-propagation process intricate and unstable, and weaken the acquisition of consistent audio-visual representations. To overcome those challenges, the stop-grad is also employed in the proposed work to decouple the gradients of the two modalities, such that the gradient of a particular modal sub-network is autonomous of the other modality. As expectation, the overall model performance can be substantially improved because each sub-network updates parameters independently. Our main contributions can be summarized as follows:

- An Audio-Visual Induction Network (AVIN) is proposed to learn a unified audio-visual representation. Based on the extracted visual feature map to generate the semantic representation of the sound source, the obtained Induction Vector can guide the network to project the audio and visual features of objects into a unified semantic space.
- The operation of stop-grad is initially introduced into the audio-visual sound source localization task to overcome the gradient coupling of distinct modalities. When the representation of one modality is regarded as constant, the gradient of the other modality can be independently obtained to decouple the gradient between the two modalities.
- To facilitate the training of the visual network, an adaptive threshold selection strategy is proposed to categorize the similarity score of the visual representation and Induction Vector into foreground, ignore, and background tri-maps, in which the optimal threshold can also be determined accordingly.
- Based on the similarity of visual features between samples as a weight, a visual weighted contrastive loss is designed for training robustness enhancement.

2 RELATED WORKS

2.1 Audio-Visual Self-Supervised Representation Learning

Audio-visual self-supervised representation learning relies on proxy tasks to generate supervised signals [50, 51]. Modern approaches solve this problem using contrastive learning. Owens et al. [29] propose a binary classification approach that considers corresponding audio-visual pairs as positive and asynchronous audio-visual pairs as negative. Korbar et al. [23] construct a two-stream network that minimizes the Euclidean distance of audio and visual features with contrastive loss, thereby ensuring that the audio-visual network is semantically coherent and temporally aligned. In contrast, Asano et al. [2] use improved Sinkhorn-Knopp algorithm [9] to assign pseudo labels to audio and visual features as supervision signals. Recently, audio-visual representation learning has been treated as an instance discrimination task [26, 27, 47, 49], in which the cosine similarity of synchronized audio-visual pairs is maximized through the use of noise contrastive estimation (NCE) or Info NCE loss. Although these works adopt a contrastive learning [28, 32] approach to learn audio-visual representations, they do not consider the issue of gradient coupling of distinct modalities. Compared to prior work, the stop-grad is employed in our network to decouple the gradients of the two modalities. By treating the representation of the corresponding modality as constant, the gradient of the current modality is solely related to itself.

2.2 Audio-Visual Sound Source Localization

Cognitive science and psychology theories suggest that visual information associated with sound would significantly enhance the searching efficiency in sound space, as demonstrated in [20]. In neuroscience, Garner et al. [13] discover that the primary visual cortex can suppress the visual responses after the association between auditory and visual stimuli. These findings have advocated the research interest in the area of audio-visual sound source localization.

Typical sound source localization relies on acoustic hardware. E.g. Zunino et al. [52] design a device equipped with a microphone array. By integrating the orientation information from the array with the visual information of the camera, the performance of visual tracking can be enhanced, but the main limitation of this scheme is the monophonic sound processing due to the complexity of required hardware. In a different way, other methods rely on spatial sparsity to determine sound source locations. Kidron et al. [21] utilizes canonical correlation analysis (CCA) to exploit the spatial sparsity of audio-visual events and avoid the issues of dimensionality. Barzelay et al. [3] employ instances of significant change within each modality to determine cross-modal associations and visual locations based on handcrafted motion cues.

With the development of deep networks, more effective techniques have been employed in recent works. Some methods [16, 29, 31, 36] exploit CAM to assist sound source localization. Based on detected object proposal boxes, [30, 37, 42, 44, 48] determine whether the potential objects are sound sources according to the learned audio feature. By computing the cosine similarity between audio-visual features, recent works [17–19, 33, 35, 39, 43] take advantage of two-stream network architecture to predict the spatial location of sound sources.

For the solutions above, aligning the data of audio and vision at the feature space is challenging because of the significant differences between the feature spaces of the two modalities, resulting in the difficulty of accurate sound source localization. Instead of aligning audio-visual modalities in feature space directly, this work also takes into account the heterogeneity of audio and vision. The proposed Induction Network performs the alignment in semantic space, which is to ensure consistent semantic and accurate localization results.

3 METHOD

Figure 2 presents the overall architecture of the proposed AVIN, and it is a two-stream model with bifurcated audio and visual modalities fused at the bottom. The AVIN is composed of: visual network, audio network, induction module (only for training), and localization module (only for inference). The corresponding operations mainly contain: using the generated visual modality features to obtain Induction Vectors, which act as intermediate vectors to connect the audio and visual modality representations. Then, an adaptive threshold selection strategy is performed with Induction Vector to learn and induct the candidate sounding object regions in the image, which is further to obtain a unified and discriminative representation in the common semantic space. Finally for the audio modality, a visual weighted contrastive loss is designed to align the audio with the visual representation, which is able to avoid faulty negative samples during training phase.

Given a video clip, the central frame $v \in \mathbb{R}^{3 \times H \times W}$ together with a 3s audio log-mel spectrogram $a \in \mathbb{R}^{1 \times F \times T}$ are input into the network, in which *H* and *W* denote the height and width of the frame, *F* represents the number of mel-frequency bins of the spectrogram, and *T* is the number of audio frames. The functionality of each part in AVIN is elaborated as below.

3.1 Visual Network

The visual network consists of a Projector P^v and a Visual Encoder E^v , which is formed by the ResNet or Transformer alternatively.

ResNet: As in [5], ResNet18 is employed as our visual encoder, which consists of 8 residual blocks [15]. In order to maintain the spatial information of the output, the average pooling layer and the fully connected layer are excluded at the end of the network. The visual embedding is denoted as $z_{RN}^v \in \mathbb{R}^{e_{RN}^v \times h \times w}$, where c_v represents the number of channels of the feature. The $h = \lfloor \frac{H}{16} \rfloor$ and $w = \lfloor \frac{W}{16} \rfloor$ denote the height and width of z_{RN}^v , respectively.

Transformer: The Vision Transformer [11] reshapes the image into a sequence of non-overlapping patch sets $v' \in \mathbb{R}^{n \times (p^2 \times 3)}$, where $p \times p$ is the resolution of each patch, and $n = HW/p^2$ represents the number of patches. By concatenating a learnable CLS token before the first patch, a total of (n + 1) tokens can be obtained. A positional embedding is then added to the token before feeding into the Transformer Encoder [45], which consists of 12 Transformer blocks. Each block is composed of a Multi-Layer Perception (MLP), Muti-Head Self-Attention (MHSA) layer, and a normalization layer. In MHSA, each token is projected into a query Q, key K, and value V, the attention between tokens is computed using:

Attention
$$(Q, K, V) = \operatorname{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$
 (1)

where d_k is the dimension of the hidden layer in MHSA. In accordance with TokenCut [46], the weights of the self-supervised trained DINO [4] are utilized to initialize the Vision Transformer. The key of the last MHSA layer in the final block is used as the visual embedding $z_{TF}^v \in \mathbb{R}^{c_{TF}^v \times h \times w}$ of the transformer, and the CLS token is discarded.

Visual Projector: To align audio and visual embeddings, a projector P^v is proposed to project the embeddings into a common space $f^v = P^v(z^v), f^v \in \mathbb{R}^{c \times h \times w}$ with dimension *c*. In our work, distinct projectors have been employed for different network architectures, with a convolutional layer (Conv) utilized for ResNet, as well as a Conv-ReLU-Conv configuration for transformer. The reason behind using different projectors for different backbones is that the parameters of ResNet are updated during training in comparison to the fixed parameters of transformer. Hence, additional nonlinear and convolutional layers can be incorporated to augment the expressive capabilities.

3.2 Audio Network

Similar to the vision network, the audio network consists of an audio encoder and an audio projector as well.



Figure 2: Architecture of Audio-Visual Induction Network. The four parts of the network: visual network, audio network, induction module and localization module are distinguished by different colors. Note that the induction module is only used during training and the localization module is only used during inference.

Audio Encoder: In this work, ResNet22 pre-trained by PANNs [22] acts as the audio encoder, which consists of 8 residual blocks, 4 supplementary convolutional layers, and two fully connected (FC) layers. To generate audio embeddings, the last classification layer is removed. The output of the audio encoder is symbolized as $z^a \in \mathbb{R}^{c^a \times 1}$, where c^a denotes the number of channels in the audio embedding.

Audio Projector: The projector is employed to project audio embeddings into a common space for the computation of similarity between audio and visual representations. Since audio embeddings are represented as one-dimensional vectors, we employ the FC-ReLU-FC structure to derive the audio representation $f^a = P^a(z^a)$, $f^a \in \mathbb{R}^{c \times 1}$, with an equal number of channels to the visual representation.

3.3 Induction Module

The audio and visual representations are connected in induction module with an intermediate vector 'Induction Vector', defined as f^{ind} , which is obtained from visual modality and is supposed to represent the semantics of the sound source. For visual networks, f^{ind} is utilized to induct f^v to acquire more precise object representations through a bootstrapping fashion. For audio networks, f^a can be aligned with f^{ind} using a visual weighted contrastive loss.

3.3.1 Induction Vector Generation.

To obtain f^{ind} from f^v , global average pooling (GAP) is performed

in visual network based on ResNet as backbone:

$$f^{ind} = \text{GAP}(f^v) \tag{2}$$

Considering that the ResNet classification network projects the pooled features linearly into logits in the category space, the pooled feature is further utilized based on the guiding intuition that contains a specific category of semantic information. For validation, the cosine similarity is calculated between f^{ind} and the visual representation map $f^{v}(i, j)$ at spatial location (i, j) using Equation 3:

$$s^{vv}(i,j) = \frac{\left\langle f^{v}(i,j), f^{ind} \right\rangle}{\|f^{v}(i,j)\|_{2} \|f^{ind}\|_{2}}, (i,j) \in [h] \times [w]$$
(3)

where $\langle \cdot, \cdot \rangle$ denotes the inner product, $s^{vv} \in \mathbb{R}^{h \times w}$. Figure 3 (a) depicts the visualization outputs of s^{vv} . The foreground region in the image has a high score, which indicates that the pooled feature f^{ind} has a strong similarity to the foreground. Based on the assumption that the existence of sounding objects in image, the target representation from pre-training contained in f^{ind} can serve as the semantics of the sound source.

For visual networks using Transformer as the backbone, an unsupervised object detection method TokenCut [46] is employed to generate f^{ind} . Based on the finding of DINO (self-distillation with no labels trained vision transformer features contain information of the semantic segmentation), TokenCut adopts Normalized Cut (NCut) [38] to divide Key features from the last attention layer of

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada.

the DINO-pretrained model into two sets, foreground and background, as shown in Figure 3 (b). To obtain the foreground, the similarity matrix M of z_v^{TF} is computed based on the Equation 4 as:

$$M_{s}(i,j) = \frac{\langle z_{i}, z_{j} \rangle}{\|z_{i}\|_{2} \|z_{j}\|_{2}}, z_{i}, z_{j} \in \{z_{v}^{TF}\}$$
(4)

where z_i and z_j represent tokens in z_{TF}^o , M is an $n \times n$ symmetrical matrix. M_b is derived after binarizing M as:

$$M_b(i,j) = \begin{cases} 1 & M(i,j) \ge \tau_m \\ \epsilon & M(i,j) < \tau_m \end{cases}$$
(5)

where $\tau_m = 0.2$ and $\epsilon = 10^{-5}$ are the hyperparameters. The set composed of z_{TF}^v is denoted as $\mathcal{Z} = \{z_i\}$, with each element z_i as the node, and the similarity between two nodes as the edge denoted by \mathcal{M} . A fully connected undirected graph $\mathcal{G} = (\mathcal{Z}, \mathcal{M})$ can be constructed with the edges and nodes, and each node is linked to others by edges. To divide the graph into two disjoint sets \mathcal{A} and \mathcal{B} , NCut performs the minimization as:

$$\frac{C(\mathcal{A},\mathcal{B})}{C(\mathcal{A},\mathcal{Z})} + \frac{C(\mathcal{A},\mathcal{B})}{C(\mathcal{B},\mathcal{Z})}$$
(6)

where $C(\cdot, \cdot)$ represents the sum of edges between the nodes within the two sets. According to [38], solving a generalized eigensystem $(D - M_b) y = \lambda Dy$ enables the discovery of the second smallest eigenvector y_1 , where *D* is:

$$D(i,j) = \begin{cases} \sum_{j} M_b(i,j) & i=j\\ 0 & i\neq j \end{cases}$$
(7)

Then the divided sets are $\mathcal{A} = \{z_i | y_1^i \leq \bar{y}_1\}$ and $\mathcal{B} = \{z_i | y_1^i > \bar{y}_1\}$, where \bar{y}_1 is the average of y_1 . The set of tokens with the largest absolute value of the eigenvalue is denoted as the foreground set \mathcal{F} , and the foreground mask $m \in \mathbb{R}^n$ is obtained by

$$m_i = \begin{cases} 1 & z_i \in \mathcal{F} \\ 0 & z_i \notin \mathcal{F} \end{cases}$$
(8)

where *m* is reshaped to $1 \times h \times w$, and the Induction Vector f^{ind} can be generated by

$$f^{ind} = \text{GAP}(f_v \circ m) \tag{9}$$

where \circ is the Hadamard product.

3.3.2 Bootstrapped Induct Visual Network.

The training of the visual network has two objectives: (1) to project the region in the visual representation f_v , which semantically corresponds to the sound source, into a unified representation distinct from other semantics, and (2) to generate a high-quality Induction Vector f^{ind} from the semantically explicit visual representation f^v .

In this work, we employ the tri-map strategy in HardWay [5] to train the visual network. The proposed AVIN differs from HardWay in two aspects: 1) AVIN uses the Induction Vector to generate a similarity map for visual representation, while HardWay utilizes the audio representation; 2) A proposed adaptive threshold selection strategy is employed to obtain the tri-map as shown in Figure 4 (a), while HardWay employs a fixed threshold.



Figure 3: Visualization of salient region. (a) Pixel-wise cosine similarity score map of ResNet features after global pooling on the feature map. (b) Foreground mask extracted by TokenCut for transformer output features.

The similarity map s^{vv} in Equation 3 is used to separate the image into positive, ignore, and negative regions, i.e., tri-map. It is pre-defined in AVIN that a certain percentage is attributed to the foreground region t_p % for each image, while t_n % is assigned to the background region, and the remaining areas are ignored. Specifically, the scores of $h \times w$ in the similarity map s^{vv} are sorted in ascending order. The minimum value of the top t_p % of the scores is denoted as ϵ_p , while the maximum value of the bottom t_n % scores is denoted as ϵ_n .

$$\hat{m}_{ijp} = \text{sigmoid} \left(\left(S_{ij} - \epsilon_p \right) / \tau_s \right)$$

$$\hat{m}_{ijn} = 1 - \text{sigmoid} \left(\left(S_{ij} - \epsilon_n \right) / \tau_s \right)$$

$$SP_{ij} = \frac{\left\langle \hat{m}_{ijp}, S_{ij} \right\rangle_F}{\left| \hat{m}_{ijp} \right|}$$

$$SN_{ij} = \frac{\left\langle \hat{m}_{ijn}, S_{ij} \right\rangle_F}{\left| \hat{m}_{ijn} \right|}$$
(10)

where \hat{m} is the soft mask, $\langle \cdot, \cdot \rangle_F$ is the Frobenius inner product, $\tau_s = 0.03$ is the hyperparameter that controls the degree of softening, and N represents the number of samples in a batch. S_{ij} denotes the cosine similarity map between the *j*-th Induction Vector f^{ind} and the visual representation map f^v of the *i*-th sample. For the training of visual network, infoNCE is adopted as the loss function:

$$\mathcal{L}_{v} = -\frac{1}{N} \sum_{i=1}^{N} \left[\log \frac{\exp\left(SP_{ii}/\tau_{c}\right)}{\sum_{j} \exp\left(SP_{ij}/\tau_{c}\right) + \sum_{j} \exp\left(SN_{ij}/\tau_{c}\right)} \right]$$
(11)

with $\tau_c = 0.07$ as the temperature hyperparameter of infoNCE.

3.3.3 Audio Representation Learning.

To establish a correlation between audio and visual representations, the Induction Vector serves as a bridge to connect the two modalities. As shown in Figure 4 (b), a visual weighted contrastive loss is



Figure 4: Schematic of (a) adaptive threshold selection strategy and (b) visual weighted contrastive loss.

introduced to facilitate learning the projection of audio representation in a common space.

$$\begin{aligned} \operatorname{Eu}_{ij} &= d \left(f_i^{ind}, f_j^a \right)^2 \\ \gamma_{ij} &= -\operatorname{cossim} \left(f_i^{ind}, f_j^{ind} \right) \\ \mathcal{L}_a &= \frac{1}{N} \sum_{i=1}^N \max \left(0, \operatorname{Eu}_{ii} - \frac{1}{N-1} \sum_{j=1}^N 1 \left(i \neq j \right) \cdot \gamma_{ij} \cdot \operatorname{Eu}_{ij} + \theta \right) \end{aligned}$$
(12)

where $d(\cdot, \cdot)$ denotes the Euclidean distance between two vectors, the margin hyperparameter $\theta = 0.6$ denotes the minimum margin between positive and negative samples. γ_{ij} represents the negative cosine similarity of the Induction Vector between two distinct samples, and it controls the direction, as well as the scale of the negative samples in the contrastive loss. It is noteworthy that f^{ind} is generated by the visual modality for the representation of object in the image. If the visual similarity between two samples is high, i.e., $\gamma < 0$, their sound is supposed to be similar as well, and it indicates that the *j*-th sample should be classified as a positive sample. On the other hand, if the visual similarity between two samples is low, i.e., $\gamma \ge 0$, it means that they are likely to be different objects and sound different either. Thus, the *j*-th and *i*-th samples are classified as negative sample pairs with the weight γ accordingly.

3.3.4 Stop-Grad Consideration.

The loss function of AVIN is defined as:

$$\mathcal{L} = \mathcal{L}_v + \mathcal{L}_a \tag{13}$$

During the training phase, the gradients of the visual and audio networks are decoupled. Specifically, for \mathcal{L}_a as given by Equation 12, all vectors associated with the visual modality, including f^{ind} and γ , are regarded as constants. A reasonable explanation is that the intricate mutual coupling of gradients may destabilize the backpropagation process, which is caused by the representation differences between the audio and visual modalities. Decoupling the gradients of the two modalities ensures that the parameter update process of the visual network would only relate to the visual modality itself. The audio network regards the Induction Vector from the visual modality as constant, and seeks to minimize the

Tianyu Liu, Peng Zhang, Wei Huang, Yufei Zha, Tao You, and Yanning Zhang

		Flickr test set		VGG-SS	
Method	Training set	cIoU	AUC	cIoU	AUC
HardWay [5]		0.582	0.525	0.288	0.351
SSPL [39]		0.743	0.587	0.208	0.300
FNAC [40]		0.843	0.633	0.336	0.372
FNAC+OGL [40]	Flickr 10k	0.847	0.643	0.407	0.404
AVIN-RN		0.868	0.659	0.423	0.421
AVIN-TF		0.843	0.632	0.413	0.418
HardWay [5]		0.699	0.573	0.269	0.344
SSPL [39]		0.759	0.610	0.289	0.356
FNAC [40]		0.787	0.593	0.348	0.380
FNAC+OGL [40]	Flickr 144k	0.840	0.631	0.406	0.403
AVIN-RN		0.872	0.658	0.423	0.420
AVIN-TF		0.843	0.639	0.422	0.419
HardWay [5]		0.618	0.536	0.291	0.368
SSPL [39]		0.763	0.591	0.316	0.374
FNAC [40]	1.000	0.857	0.637	0.372	0.388
FNAC+OGL [40]	VGGSound 10k	0.821	0.636	0.415	0.408
AVIN-RN		0.884	0.659	0.450	0.431
AVIN-TF		0.835	0.643	0.448	0.433
HardWay [5]		0.719	0.582	0.292	0.367
SSPL [39]		0.767	0.605	0.323	0.376
FNAC [40]	Magan Land	0.847	0.638	0.406	0.405
FNAC+OGL [40]	VGGSound 144k	0.851	0.643	0.421	0.412
AVIN-RN		0.876	0.658	0.449	0.436
AVIN-TF		0.851	0.644	0.448	0.433

Table 1: Sound source localization result on Flickr test set and VGG-SS benchmark

Euclidean distance between the synchronized audio representation and the Induction Vector.

3.4 Localization Module

During the inference stage, the localization module is to ascertain the similarity between the audio representation and the visual representation map. Given the audio representation f^a and the visual representation map f^v in the learned common space, the similarity map s^{av} is:

$$s^{av}(i,j) = \frac{\langle f^v(i,j), f^a \rangle}{\|f^v(i,j)\|_2 \|f^a\|_2}, (i,j) \in [h] \times [w]$$
(14)

Finally, a min-max normalization process is performed to re-scale s^{av} to the interval [0, 1]:

$$\tilde{s}^{av} = \frac{s^{av} - \min(s^{av})}{\max(s^{av}) - \min(s^{av})}$$
(15)

where \tilde{s}^{av} represents the output of AVIN and denotes the degree of correlation between the location of each image and the provided audio cues.

4 EXPERIMENTS

4.1 Comparisons with State-of-the-art Methods

The proposed AVIN is firstly compared with other works on the SoundNet-Flickr test set as shown in Table 1. We employ ResNet and Transformer as visual encoders denoted as AVIN-RN and AVIN-TF, respectively. When training on Flickr 10k and 144k, AVIN-TF is

comparable to the recently proposed FNAC [40], while AVIN-RN outperforms the previous best (0.868 vs. 0.847 in 10k and 0.872 vs. 0.840 in 144k). Noticed that FNAC+OGL incorporates Object-Guided Localization (OGL), which is a post-processing strategy to refine localization results. In comparison, the output of AVIN is only based on the correspondence of audio and visual features, but achieves superior results. For cross-dataset evaluation purposes, the proposed models are trained on the VGGSound 10k and 144k training sets. Since a greater diversity of video categories is presented in VGGSound compared to SoundNet-Flickr, the AVIN can effectively establish the association between visual and audio using the induction vector, which has achieved state-of-the-art results and validated the cross-dataset generalization ability in both settings.

The evaluation results on VGG-SS benchmark are also illustrated in Table 1. With the multiple reproductions, the best results are reported because the sample count in the test set is less than [5] (4664 vs. 5158). AVIN surpasses all the other works with a clear margin. In the VGGSound 10k training case, AVIN-RN outperforms FNAC+OGL by 8.4% cIoU and 5.6% AUC performance increase, with cIoU of 0.450 and AUC of 0.431. All the results demonstrate the superior performance of the proposed work compared to state-ofthe-art works on both datasets.

Furthermore, we observe that smaller subsets (Flickr or VG-GSound 10k) exhibit comparable performance to larger ones (Flickr or VGGSound 144k). We hypothesize that AVIN can learn sufficient audio-visual semantic information for satisfactory fitting from a smaller subset. In contrast, the larger subset not only provides less additional semantic information based on the smaller subset but also may affected by overfitting [25], as evidenced in the AVIN-RN model's performance on the VGGSound dataset in Table 1.

The visualized results for sound source localization of AVIN-RN and AVIN-TF on Flickr test set and VGG-SS are shown in Appendix A.5 Figure 5. Our AVIN demonstrates enhanced prediction in localizing the semantic region of the sound source compared to prior works while minimizing background interference. Notably, AVIN-TF exhibits a certain capability to delineate the contours of sound sources.

4.2 Ablation Study

In the ablation study conducted in the subsequent experiments, Flickr refers to using SoundNet-Flickr 10k as the training set and evaluating on the SoundNet-Flickr test set, while VGG-SS refers to using VGGSound 10k as the training set and evaluating on VGG-SS.

4.2.1 Induction Vector.

To evaluate the contribution of the induction vector, two sets of experiments are conducted: (a) remove the induction vector and visual weighted contrastive loss, directly compute the cosine similarity between the visual representation map f^v and the audio vector f^a , which means to use the output of the localization module s^{av} instead of s^{vv} to generate the tri-map; (b) retain f^{ind} but remove \mathcal{L}_v . Table 2 shows the results of the experiments above. In case (a), when f^{ind} is removed, the cloU performance of AVIN-TF drops to 0.337 (on Flickr) and 0.098(on VGG-SS), while AVIN-RN drops to 0.659(on Flickr) and 0.361(on VGG-SS). A reasonable explanation is that the location of salient objects contained in the pre-trained

Table 2: Ablation study for the induction v	ecto
---	------

	Method	Dataset	cIoU	AUC
(a)	AVIN-RN AVIN-TF	Flickr VGGSound Flickr VGGSound	0.659 0.361 0.337 0.098	0.560 0.395 0.436 0.251
	AVIN-RN	Flickr	0.522	0.491
(b)	AVIN-TF	Flickr VGGSound	0.542 0.255	0.505 0.338

ResNet features can facilitate the network to determine rough audiovisual correspondence, and in accompany with the induction vector to benefit AVIN-RN to learn precise correspondence. Due to the lack of salient object information, the result of the AVIN-TF is poor after removing the induction vector. In case (b), the performance of AVIN-RN and AVIN-TF drops to similar levels because of insufficient training of bootstrapped visual network, which further validates an informative deficiency without training visual network.

4.2.2 Stop-Grad Operation.

Experiments are conducted to investigate the influence of stop-grad operation on training as shown in Table 3. It can be found that AVIN-RN is more sensitive to the gradient than AVIN-TF, and the cIoU performance drops by 0.559 on Flickr and 0.273 on VGG-SS, while AVIN-TF has a drop of 0.092 on Flickr and 0.081 on VGG-SS. The reason is that the gradient updates of the visual and audio networks mainly depend on both modalities' representations simultaneously, which may weaken the training process due to the isomerism of modality. By allowing the visual/audio network to update its parameters only according to the gradient of the visual/audio modality, the stop-grad operation can help to improve the performance of audio-visual sound source localization.

To verify the effect of stop-grad on other audio-visual sound source localization architectures, a HardWay [5] variant is used with the adaptive threshold selection strategy. The experiments are conducted by replacing the audio network with a pretrained ResNet22 model with fixed parameters [22]. The visual projector uses a convolutional layer and the audio projector uses a fully connected (FC) layer with the same number of channels. An FNAC variant is also conducted by replacing the audio network with fixed ResNet22. The experimental results are shown in Table 4, which have verified that the performance is still acceptable with stop-grad even the audio network parameters are fixed. When the audio representation has a gradient and the audio network parameters are updated with training, the performance decreases significantly in comparison to stop-grad. The results indicates that the stop-grad is an effective plug-and-play operation to improve the performance of other architectures, and also promising to benefit the design of future audio-visual networks.

4.2.3 Visual Weighted Contrastive Loss.

To validate the robustness of visual weights for AVIN, the visual weight parameter γ is set to 1 with the vanilla contrastive loss to

Table 3: Ablation study for stop-grad and visual weightedcontrastive loss

	Weighted	Method	Flickr		VGGSound	
Stop-grad			cIoU	AUC	cIoU	AUC
×	\checkmark	AVIN-RN	0.309	0.411	0.177	0.300
		AVIN-TF	0.751	0.578	0.367	0.397
\checkmark	×	AVIN-RN	0.851	0.650	0.436	0.431
	~	AVIN-TF	0.755	0.602	0.437	0.419
\checkmark	/	AVIN-RN	0.868	0.659	0.450	0.431
	\checkmark	AVIN-TF	0.843	0.632	0.448	0.433

Table 4: Stop-grad on HardWay [5] and FNAC [40] variant

		Flickr		VGGSound	
Method	stop-grad	cIoU	AUC	cIoU	AUC
Hardway [5]	×	0.731	0.583	0.401	0.414
	✓	0.811	0.614	0.412	0.419
FNAC [40]	×	0.755	0.588	0.368	0.389
	✓	0.779	0.608	0.372	0.390

train the audio network. The results presented in Table 3 (b) indicate that the performance of AVIN for both two architectures drops to different degrees. Since the vanilla contrastive loss cannot distinguish between audio-visual pairs with the same semantics contained in negative pairs, the Euclidean distance between f^{ind} and f^a is incorrectly maximized. Comparatively, the proposed visual weighted contrastive loss, which is based on visual priors, can correct and weight the erroneous negative pairs as positive pairs, thereby substantially enhance the overall robustness.

4.2.4 Percentage of Adaptive Threshold.

The performance of network is affected by the hyperparameter of adaptive threshold selection strategy. As a solution, we train AVIN-RN and AVIN-TF with different t_p and t_n combinations respectively, and the results are shown in Table 5. For AVIN-RN, the results show less susceptibility to the ratios, and a broad range of thresholding ratios (20-40% for t_p , 10-50% for t_n) yield satisfactory outcomes. However, AVIN-TF is more sensitive to t_p and can achieve better results when t_p is 20-30%.

Additionally, we replace the tri-map with the bi-map generated by TokenCut, which involves reshaping the mask *m* in Equation 8 to the scale of $h \times w$, where $\hat{m}_{iip} = m$, $\hat{m}_{iin} = 1 - m$, $\hat{m}_{ijn} = 1$ in Equation 10, and 1 denotes a $h \times w$ tensor of all ones. Compared to the performance of bi-map in last row of Table 5, the tri-map generated by the adaptive threshold selection strategy can achieve better performance.

5 CONCLUSION

In this work, Induction Network is proposed to bridge the gap between audio and visual modalities. After decoupling the gradients of different modalities, the audio and visual representations are aligned by the Induction Vector, which is obtained from the visual Tianyu Liu, Peng Zhang, Wei Huang, Yufei Zha, Tao You, and Yanning Zhang

Table 5: Ablation study for adaptive threshold

			Flickr		VGGSound	
Method	t _p (%)	t_n (%)	cIoU	AUC	cIoU	AUC
	10	30	0.843	0.637	0.414	0.420
	10	50	0.847	0.636	0.417	0.420
	20	10	0.868	0.650	0.429	0.426
	20	30	0.872	0.651	0.431	0.427
AVIN-RN	20	50	0.859	0.648	0.440	0.431
	30	30	0.868	0.658	0.446	0.430
	30	50	0.868	0.659	0.450	0.431
	40	30	0.868	0.666	0.428	0.423
	40	50	0.868	0.666	0.431	0.423
	10	30	0.562	0.506	0.413	0.420
AVIN-TF	10	50	0.546	0.502	0.419	0.422
	20	10	0.791	0.607	0.463	0.441
	20	30	0.791	0.598	0.478	0.449
	20	50	0.795	0.604	0.478	0.447
	30	30	0.843	0.632	0.448	0.433
	30	50	0.827	0.634	0.446	0.433
	40	30	0.807	0.639	0.409	0.413
	40	50	0.819	0.643	0.392	0.406
	-	-	0.755	0.594	0.459	0.441

modality in a bootstrap manner. Nevertheless, an adaptive threshold selection strategy and a visually weighted contrastive loss are proposed to further improve the robustness of the network.

Limitations. Although visual weighted contrastive loss is used to correct faulty negatives, the faulty positives in the training set, i.e., audio-visual irrelevant pairs, also limit the localization performance. A faulty positive mining approach is considered to mitigate this issue.

ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China (No. 61971352, No. 62271239), Ningbo Natural Science Foundation (No. 2021J048, No. 2021J049), Jiangxi Double Thousand Plan (No. JXSQ2023201022), Fundamental Research Funds for the Central Universities (No. D5000220190), Innovative Research Foundation of Ship General Performance (No. 25522108).

REFERENCES

- Relja Arandjelovic and Andrew Zisserman. 2017. Look, listen and learn. In Proceedings of the IEEE International Conference on Computer Vision. 609–617.
- [2] Yuki Asano, Mandela Patrick, Christian Rupprecht, and Andrea Vedaldi. 2020. Labelling unlabelled videos from scratch with multi-modal self-supervision. Advances in Neural Information Processing Systems 33 (2020), 4660–4671.
- [3] Zohar Barzelay and Yoav Y Schechner. 2007. Harmony in motion. In 2007 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 1–8.
- [4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. In Proceedings of the IEEE/CVF international conference on computer vision. 9650–9660.
- [5] Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. 2021. Localizing visual sounds the hard way. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 16867–16876.
- [6] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. 2020. Vggsound: A large-scale audio-visual dataset. In ICASSP 2020-2020 IEEE International

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada.

Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 721–725.

- [7] Xinlei Chen and Kaiming He. 2021. Exploring simple siamese representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 15750–15758.
- [8] C.K. Cowley and D.M. Jones. 1992. More than meets the eye: issues relating to the application of speech displays in human-computer interaction. *Displays* 13, 2 (1992), 69–74. https://doi.org/10.1016/0141-9382(92)90100-6
- [9] Marco Cuturi. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. Advances in neural information processing systems 26 (2013).
- [10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition. Ieee, 248–255.
- [11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020).
- [12] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein. 2018. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. arXiv preprint arXiv:1804.03619 (2018).
- [13] Aleena R Garner and Georg B Keller. 2022. A cortical circuit for audio-visual predictions. *Nature neuroscience* 25, 1 (2022), 98–105.
- [14] 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. 2020. Bootstrap your own latent-a new approach to self-supervised learning. Advances in neural information processing systems 33 (2020), 21271–21284.
- [15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition. 770–778.
- [16] Di Hu, Xuhong Li, Lichao Mou, Pu Jin, Dong Chen, Liping Jing, Xiaoxiang Zhu, and Dejing Dou. 2020. Cross-task transfer for geotagged audiovisual aerial scene recognition. In European Conference on Computer Vision. Springer, 68–84.
- [17] Di Hu, Feiping Nie, and Xuelong Li. 2019. Deep multimodal clustering for unsupervised audiovisual learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 9248–9257.
- [18] Di Hu, Rui Qian, Minyue Jiang, Xiao Tan, Shilei Wen, Errui Ding, Weiyao Lin, and Dejing Dou. 2020. Discriminative sounding objects localization via selfsupervised audiovisual matching. Advances in Neural Information Processing Systems 33 (2020), 10077–10087.
- [19] Xixi Hu, Ziyang Chen, and Andrew Owens. 2022. Mix and localize: Localizing sound sources in mixtures. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 10483–10492.
- [20] Bill Jones and Boris Kabanoff. 1975. Eye movements in auditory space perception. Perception & Psychophysics 17 (1975), 241–245.
- [21] Einat Kidron, Yoav Y Schechner, and Michael Elad. 2005. Pixels that sound. In 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), Vol. 1. IEEE, 88–95.
- [22] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley. 2020. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (2020), 2880–2894.
- [23] Bruno Korbar, Du Tran, and Lorenzo Torresani. 2018. Cooperative learning of audio and video models from self-supervised synchronization. Advances in Neural Information Processing Systems 31 (2018).
- [24] Pingchuan Ma, Stavros Petridis, and Maja Pantic. 2021. End-to-end audio-visual speech recognition with conformers. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 7613–7617.
- [25] Shentong Mo and Pedro Morgado. 2022. A Closer Look at Weakly-Supervised Audio-Visual Source Localization. In Advances in Neural Information Processing Systems, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), Vol. 35. Curran Associates, Inc., 37524–37536. https://proceedings.neurips.cc/paper_files/paper/2022/file/ f3f2ff9579ba6deeb89caa2fe1f0b99c-Paper-Conference.pdf
- [26] Pedro Morgado, Ishan Misra, and Nuno Vasconcelos. 2021. Robust audio-visual instance discrimination. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 12934–12945.
- [27] Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. 2021. Audio-visual instance discrimination with cross-modal agreement. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 12475–12486.
- [28] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018).
- [29] Andrew Owens and Alexei A Efros. 2018. Audio-visual scene analysis with self-supervised multisensory features. In Proceedings of the European Conference on Computer Vision (ECCV). 631–648.
- [30] Sanjeel Parekh, Alexey Ozerov, Slim Essid, Ngoc QK Duong, Patrick Pérez, and Gaël Richard. 2019. Identify, locate and separate: Audio-visual object extraction in large video collections using weak supervision. In 2019 IEEE Workshop on

Applications of Signal Processing to Audio and Acoustics (WASPAA). IEEE, 268–272.

- [31] Rui Qian, Di Hu, Heinrich Dinkel, Mengyue Wu, Ning Xu, and Weiyao Lin. 2020. Multiple sound sources localization from coarse to fine. In *Computer Vision–ECCV* 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16. Springer, 292–308.
- [32] Jianing Quan, Baozhen Ge, and Lei Chen. 2022. Cross attention redistribution with contrastive learning for few shot object detection. *Displays* 72 (2022), 102162. https://doi.org/10.1016/j.displa.2022.102162
- [33] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. 2018. Learning to localize sound source in visual scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 4358–4366.
- [34] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. 2019. Learning to localize sound sources in visual scenes: Analysis and applications. *IEEE transactions on pattern analysis and machine intelligence* 43, 5 (2019), 1605–1619.
- [35] Arda Senocak, Hyeonggon Ryu, Junsik Kim, and In So Kweon. 2022. Less can be more: Sound source localization with a classification model. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 3308–3317.
- [36] Rahul Sharma, Krishna Somandepalli, and Shrikanth Narayanan. 2020. Cross modal video representations for weakly supervised active speaker localization. arXiv e-prints (2020), arXiv-2003.
- [37] Jiayin Shi and Chao Ma. 2022. Unsupervised Sounding Object Localization with Bottom-Up and Top-Down Attention. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 1737–1746.
- [38] Jianbo Shi and Jitendra Malik. 2000. Normalized cuts and image segmentation. IEEE Transactions on pattern analysis and machine intelligence 22, 8 (2000), 888– 905.
- [39] Zengjie Song, Yuxi Wang, Junsong Fan, Tieniu Tan, and Zhaoxiang Zhang. 2022. Self-Supervised Predictive Learning: A Negative-Free Method for Sound Source Localization in Visual Scenes. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 3222–3231.
- [40] Weixuan Sun, Jiayi Zhang, Jianyuan Wang, Zheyuan Liu, Yiran Zhong, Tianpeng Feng, Yandong Guo, Yanhao Zhang, and Nick Barnes. 2023. Learning Audio-Visual Source Localization via False Negative Aware Contrastive Learning. arXiv preprint arXiv:2303.11302 (2023).
- [41] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2016. YFCC100M: The new data in multimedia research. *Commun. ACM* 59, 2 (2016), 64–73.
- [42] Yapeng Tian, Di Hu, and Chenliang Xu. 2021. Cyclic co-learning of sounding object visual grounding and sound separation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2745–2754.
- [43] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. 2018. Audiovisual event localization in unconstrained videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 247–263.
- [44] Thanh-Dat Truong, Chi Nhan Duong, Hoang Anh Pham, Bhiksha Raj, Ngan Le, Khoa Luu, et al. 2021. The right to talk: An audio-visual transformer approach. In Proceedings of the IEEE/CVF International Conference on Computer Vision. 1105– 1114.
- [45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. Advances in neural information processing systems 30 (2017).
- [46] Yangtao Wang, Xi Shen, Shell Xu Hu, Yuan Yuan, James L. Crowley, and Dominique Vaufreydaz. 2022. Self-Supervised Transformers for Unsupervised Object Discovery Using Normalized Cut. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 14543–14553.
- [47] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. 2018. Unsupervised feature learning via non-parametric instance discrimination. In Proceedings of the IEEE conference on computer vision and pattern recognition. 3733–3742.
- [48] Hanyu Xuan, Zhiliang Wu, Jian Yang, Yan Yan, and Xavier Alameda-Pineda. 2022. A Proposal-based Paradigm for Self-supervised Sound Source Localization in Videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 1029–1038.
- [49] Mang Ye, Xu Zhang, Pong C Yuen, and Shih-Fu Chang. 2019. Unsupervised embedding learning via invariant and spreading instance feature. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 6210– 6219.
- [50] Pengcheng Zhang, Lei Zhou, Xiao Bai, Chen Wang, Jun Zhou, Liang Zhang, and Jin Zheng. 2022. Learning multi-view visual correspondences with selfsupervision. *Displays* 72 (2022), 102160. https://doi.org/10.1016/j.displa.2022. 102160
- [51] Hao Zhu, Man-Di Luo, Rui Wang, Ai-Hua Zheng, and Ran He. 2021. Deep audiovisual learning: A survey. *International Journal of Automation and Computing* 18, 3 (2021), 351–376.
- [52] Andrea Zunino, Marco Crocco, Samuele Martelli, Andrea Trucco, Alessio Del Bue, and Vittorio Murino. 2015. Seeing the sound: A new multimodal imaging device for computer vision. In Proceedings of the IEEE International Conference on Computer Vision Workshops. 6–14.

Tianyu Liu, Peng Zhang, Wei Huang, Yufei Zha, Tao You, and Yanning Zhang

A APPENDIX

A.1 Datasets

SoundNet-Flickr: SoundNet-Flickr [41] is a dataset consisting of over 2 million real-life image-sound pairs with 500 annotated bounding boxes by Senocak et al. [34], and each pair is processed by 3 annotators. Following [39], the training set contains a random subset of 10k and 144k pairs, while the testing set contains 250 annotated pairs.

VGG-Sound and VGG-Sound Source: VGG-Sound [6] is a more challenging dataset that includes over 200k in-the-wild video clips from YouTube with 10s audio and video segments. VGG-Sound Source (VGG-SS) is an audio-visual localization benchmark with 5k bounding box annotations of manually verified sounding objects [5]. Following [39], the training set for both datasets includes a random subset of 10k and 144k pairs, while the testing set contains a subset (4664 samples) of VGG-SS due to some unavailable videos.

A.2 Implementation Details

Visual Network: The ResNet18 [15] and ViT-S [11] pretrained on ImageNet [10] are employed as visual encoder. ResNet18 comprises 8 residual blocks and the output channel is 512, ViT-S contains 12 blocks with the 6 heads and 384 channel in MHSA. The length of the image patch is 16. The parameters of ResNet18 are updated while those of ViT-S remain frozen during training. The visual projector projects the output channel to 512 for both ResNet and Transformer

Audio Network: The ResNet22 model pretrained by PANNs [22] is employed as the audio encoder, including 8 residual blocks and the channel number of output embedding is 2048. The parameters of the audio encoder are frozen during training, and the output of the audio projector is a 512D vector.

Data: For data requirement, the middle frame of video clip and the 3-second sound surrounding the frame are selected as the visual and audio input, respectively. We performed operations of 224×224 random cropping and random horizontal flipping on the input images for data augmentation during training. Following [22], the audio signal is resampled to 32kHz, and STFT is applied on waveforms with a Hamming window size of 1024 and a hop size of 320. The log-mel spectrogram is computed by applying 64 mel-filter banks, which is transformed to 301×64 .

Training Details: The Adam optimizer is used with the rate of 1×10^{-5} for the ResNet18 backbone, as well as the learning rate of 1×10^{-4} for both the visual and audio projectors. The a batch size is set to 256 for training, and early stopping is configured to avoid overfitting. The hyperparameter of adaptive threshold selection strategy is set to $t_p = 30$ and $t_n = 50$ for AVIN-RN, $t_p = 30$ and $t_n = 30$ for AVIN-TF.

A.3 Evaluation Metrics

Following [34][5][39], the metrics of cIoU and AUC are used for performance evaluation. The score map g is computed for each sample, which is defined as:

$$g = \min\left(\sum_{j=1}^{n} \frac{b_j}{C}, 1\right)$$
(16)

where b_j is the binary image of *j*-th bounding box, *C* is the minimum number of opinions to reach an agreement and C = 2 in practice, thus cloU is defined as:

$$cIoU(t) = \frac{\sum_{i \in A(t)} g_i}{\sum_i g_i + \sum_{i \in A(t) - G} 1}$$
(17)

where *i* is the pixel index of the score map, and the decision threshold *t* is set to 0.5. $A(t) = \{i|s_i > t\}$ and $G = \{i|g_i > 0\}$, where s_i indicates the activation of heatmap *S* at location *i*. AUC is the area under the curve plotted by the ratio of samples with cIoU > t' to the total number of samples when t' changing from 0 to 1.

A.4 Computational Complexity Analysis

We perform the computational complexity analysis for different methods as shown in Table 6. The number of parameters (column 5) for AVIN-RN is 76.4M, whereas AVIN-TF is 89.5M (where ResNet22 [22] contributes 63.6M). It is worth noting that the SSPL has more parameters than the proposed model (108.6M vs. 76.4M/89.5M) but exhibits lower performance.

Additionally, the floating point operations (FLOPs) for each method (column 4) is also calculated, and the AVIN is able to maintain a modest computational cost as well as to achieve a better performance simultaneously.

Finally, we compare the speed of training (column 2) and inference (column 3) for all the methods. The training process is running on two 2080Ti GPUs, whereas inference is on a single 2080Ti GPU. It is normal for certain models to exhibit faster training speed compared to inference, which is owing to the acceleration of two GPUs.

Among the proposed models, AVIN-RN demonstrates the fastest training speed, which can process approximately 580 audio-visual pairs per second. AVIN-TF exhibits a relatively slower training speed, approximately 156 pairs per second, but still faster than SSPL. This discrepancy is affected by the TokenCut [46] that runs on the CPU and lacks acceleration, and thus consequently requires a longer duration. In the inference stage, AVIN-RN maintains its superiority in terms of speed, and achieves around 560 audio-visual pairs per second. Comparatively, AVIN-TF achieves 385 pairs per second, which is similar as FNAC and faster than SSPL.

A.5 Visualization results

In this section, visualization results of HardWay [5], SSPL [39], FNAC [40], FNAC+OGL, AVIN-RN and AVIN-TF are shown in Figure 5.

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada.

Table 6: Computational complexity analysis of different methods

Method	Train (AVpairs/s)	Inference (AVpairs/s)	FLOPs	Param
HardWay [5]	474	430	5.5G	23.4M
SSPL [39]	90	225	46.7G	108.6M
FNAC [40]	492	394	5.6G	22.9M
FNAC+OGL [40]	492	351	7.4G	34M
AVIN-RN	580	560	10.6G	76.4M
AVIN-TF	156	385	12.3G	89.5M



Figure 5: Visualization of different methods on Flickr test set and VGG-SS benchmark. Both AVIN-RN and AVIN-TF are able to localize sound sources in a variety of challenging scenarios.