Acoustic and Visual Knowledge Distillation for Contrastive Audio-Visual Localization

Ehsan Yaghoubi
ehsan.yaghoubi@uni-hamburg.de
Universität Hamburg
Hamburg, Germany

André Kelm
andre.kelm@uni-hamburg.de
Universität Hamburg
Hamburg, Germany

Timo Gerkmann
timo.gerkmann@uni-hamburg.de
Universität Hamburg
Hamburg, Germany

Simone Frintrop
simone.frintrop@uni-hamburg.de
Universität Hamburg
Hamburg, Germany

ABSTRACT

This paper introduces an unsupervised model for audio-visual localization, which aims to identify regions in the visual data that produce sounds. Our key technical contribution is to demonstrate that using distilled prior knowledge of both sounds and objects in an unsupervised learning phase can improve performance significantly. We propose an Audio-Visual Correspondence (AVC) model consisting of an audio and a vision student, which are respectively supervised by an audio teacher (audio recognition model) and a vision teacher (object detection model). Leveraging a contrastive learning approach, the AVC student model extracts features from sounds and images and computes a localization map, discovering the regions of the visual data that correspond to the sound signal. Simultaneously, the teacher models provide feature-based hints from their last layers to supervise the AVC model in the training phase. In the test phase, the teachers are removed. Our extensive experiments show that the proposed model outperforms the state-of-the-art audio-visual localization models on 10k and 144k subsets of the Flickr and VGGS datasets, including cross-dataset validation.

KEYWORDS

Audio-visual representation learning, Knowledge distillation, cross-modal learning, sound-image localization, multi-modal teacher-student, acoustic-visual learning.

ACM Reference Format:
Ehsan Yaghoubi, André Kelm, Timo Gerkmann, and Simone Frintrop. 2023. Acoustic and Visual Knowledge Distillation for Contrastive Audio-Visual Localization. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '23), October 09–13, 2023, Paris, France.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3577190.3614144

ICMI '23, October 09–13, 2023, Paris, France

ACM ISBN 979-8-4007-0055-2/23/10…$15.00
https://doi.org/10.1145/3577190.3614144

1 INTRODUCTION

Audio-visual localization identifies spatial regions in visual data that correlate with sound sources. This task has applications in various fields such as security [33], assistive technology [17], and speech recognition [5]. For example, assistive electronic devices with audio-visual localization can provide spatial audio and visual cues to help people with visual or hearing impairments communicate.

A promising call of existing deep learning models [18, 20, 30, 32, 39] for audio-visual localization tend to use large amounts of unlabeled videos, rather than consuming resources to provide annotation for small sets of data. One challenge in unsupervised approaches is that learning from unlabeled data usually results in worse performance than when the data is labeled. For example, [28] demonstrated that a supervised model on 2.5k labeled data achieved a cIoU of 80.4 and an AUC of 60.3; however, when the same model was trained on 144k unlabeled samples, its performance dropped to 66.0 and 55.8, respectively. To address this challenge, a general solution is to learn from audio and visual correlation in a self-supervised manner [2, 18, 21, 32, 36]. Although these methods have demonstrated potential in unsupervised audio-visual localization, their primary focus lies in association of spatial locations in visual data to the audio signal, merely using a contrastive learning strategy. However, this approach can result in either inadequate or excessive association between the visual area and the sound signal [21]. To address this issue, several solutions have been proposed [6, 21, 28, 30]. For instance, [28] suggests incorporating a portion of labeled data during the training phase, while [6] proposes a learning scheme that considers non-object backgrounds as distracting areas and tries to focus on the objects, and [21] presents an object-guided model during the inference phase to emphasize the objects in the scene. These methods learn from potential locations [21, 30, 39] or filter out the distracting regions [6], without fully considering the knowledge of acoustics and objects simultaneously. Although these methods are effective, they still rely on some labeled data, or need object guided models in the inference phase.

Similar to [21], we use contrastive learning, a well-suited approach for capturing image-sound similarities in a joint embedding space. However, to address unsuitable association between the visual area and the sound signal, we propose a knowledge distillation framework, in which we transfer the knowledge from pre-trained audio classifiers and object detectors to

our audio and visual student models, respectively. We demonstrate that this approach is simple but effective as it does not require training the teacher models nor add complexity to the inference phase. Evaluations on well-known datasets confirm the effectiveness of our approach.

Our main contributions are:

- we introduce an unsupervised student-teacher architecture for audio-visual localization that leverages distilled prior knowledge of objects and sounds to improve the performance of a simple Audio-Visual Correspondence (AVC) model, without the need for labeled data.
- we introduce a knowledge distillation approach in which the students and teachers were trained on distinct *datasets* and *tasks*. This means our approach allows us to distill and transfer knowledge from one task to another, even when the datasets employed in these tasks vary in terms of type, annotations, and size.
- we demonstrate the effectiveness of both visual- and audio-based prior knowledge for audio-visual localization, addressing a gap in the existing research domain.
- we show that our proposed method achieves state-of-the-art results in unsupervised learning scenarios and cross-dataset validation on the Flickr [3] and VGGS datasets [7].

2 RELATED WORK

Our proposed method focuses on audio-visual localization and falls within the scope of several research domains, including audio-visual localization, knowledge distillation and contrastive learning. In the following, we survey several papers that are related to our work in terms of their research focus.

2.1 Audio-Visual Localization

Audio-visual localization has been extensively studied in the literature [1, 6, 7, 18, 20–23, 25, 28–30, 32, 36, 39]. In general, the audio-visual localization models consist of three building blocks: A sound encoder, a visual encoder, and a localization block [40]. The sound and visual building blocks extract generalized features from audio and visual data, respectively, while the last block computes an audio-visual localization map that pinpoints the spatial location of the acoustic signals on the visual data.

Since it is expensive to prepare large labeled learning benchmarks, major unlabeled audio-visual localization datasets [3, 7] were collected from Flickr and Youtube, and a small part of them was later annotated for evaluation of the algorithms [6, 28]. These datasets are video-based; however, instead of analyzing several frames [27], many methods [6, 21, 28] extract one frame from the visual data and use a few seconds of the acoustic data around it to perform the audio-visual localization task. For example, [21, 28] extract the middle frame from the video clips and use the middle 3 seconds of the sound and process it in the frequency domain, while [30] takes the middle frame and uses the first 20 seconds of the sound signal.

Another aspect that is common in the literature is the use of unsupervised and self-supervised learning strategies [18, 30, 32]. For instance, [18] presents an iterative contrastive representation learning approach, in which, instead of using prior assumptions or labels, it leverages the relationships between audio and visual data in one epoch for the next epochs and additionally suggests finding the positive samples for learning a contrastive representation from different video clips. [32] is a recent work that proposes a self-supervised method to improve the contrastive learning strategy by providing reliable positive and negative samples.

It is worth mentioning that some existing works, such as [21, 25, 28] process the acoustic data as a whole, instead of separating sounds from each other [36]. For instance, [25] introduces a two-stage framework to first classify the input data and then uses class activation maps and audio-visual correspondence to relate different sounds to relevant locations on the visual data. Differently, [36] uses a sound separation sub-network to separate sound sources and locate the relevant visual source in the visual data. This idea can help to filter out the sounds from non-objects and the objects that are presented in the scene but do not emit any sound.

In audio-visual localization, the models localize the sounds from presented objects in the scene, rather than non-objects like the wind or absent objects [36, 39]. For example, [39] uses a visual reasoning module to first recognize the regions with objects, and then uses the correlation between audio and visual data to guide the model to recognize which objects in the scene produce the sound.

In conclusion, while the current literature on audio-visual localization has made notable strides in precisely localizing sounds within visual scenes, several challenges remain, which we aim to tackle in this paper. These include examining the impact of acoustic and visual knowledge on model performance, utilizing large pre-trained models during the learning phase rather than the inference phase, and addressing the issue of low accuracy in unsupervised audio-visual localization tasks.

2.2 Knowledge distillation

Knowledge distillation is used to transfer knowledge from a large, complex model (the teacher) to a smaller, simpler model (the student) [12]. The student model is able to learn from the teacher's logits or predictions, known as hints, without having to process all of the data that the teacher did. This may result in similar performance as the teacher, while using fewer resources.

In the context of cross-modal knowledge distillation, Perez *et al.* [24] introduce a model for audio-visual action recognition, in which one acoustic-student model learns from two teachers trained on heterogeneous data, i.e., RGB video, raw audio signals, and acoustic images, in which sound waves are represented in a three-dimensional space. In addition, Chen *et al.* [9] propose a framework for video-based human activity recognition, in which one 3D video-based student model learns from two pre-trained teachers. Compared to prior studies [9, 24], not only we consider different research questions and datasets, but also we propose a knowledge distillation architecture that comprises two students and two teachers. Specifically, we employ an audio teacher to supervise the audio student, and a separate visual teacher to oversee the visual student. This approach is distinct from previous works [9, 24], which often utilize only one student to learn both modalities.

2.3 Contrastive learning for cross-modal data

Contrastive learning is categorized as an unsupervised learning approach that aims to learn a representation of data that clusters the positive (similar) samples together and pushes the negative (dissimilar) instances apart. Prominent works such as [8], [13], and [19] showed that a proper contrastive learning regime on the visual-based datasets can significantly close the performance gap between unsupervised and supervised approaches, which encouraged more exploration of contrastive learning techniques for cross-modal tasks. For example, CLIP [26] is a contrastive learning framework that can learn joint representations of images and text. The CLIP model is trained on a massive dataset of image-text pairs, which allows it to learn to associate images with relevant text descriptions and vice versa. In this context, [1, 2, 14, 21, 25, 28] use contrastive learning to optimize the audio-visual localization map. For instance, [6] uses a noise-contrastive term, which helps to reduce the impact of trivial similarities and increase the emphasis on meaningful similarities. This is achieved by comparing each input representation with a set of negative instances, which are drawn randomly from a noise distribution. The model is then trained to maximize the difference between the similarity scores of the positive pairs and the negative pairs. We used the suggested contrastive learning approach from [21] in our AVC model, as it has demonstrated effectiveness in audio-visual localization

3 PROPOSED METHOD

3.1 Problem setting

Let $\{(a_i, v_j)|i, j = 1, \ldots, n\}$ be a dataset of n pairs of sound a_i and image v_j. Inputs a_i and v_j are correlated when $i = j$. Pre-processing details of the inputs are presented in Section 3.6. The goal is to learn an unsupervised model $F_{av}((a_i, v_j))$ that outputs a localization map $\left\{o_{ij}^{xy}|i, j = 1, \ldots, n\right\}_{i=j}$ with the same spatial size of v_j, such that xy refers to (x, y) coordinates in image v_j.

3.2 Overview

As shown in Fig. 1, our proposed audio-visual localization model comprises three components: the Audio-Visual Correspondence (AVC) model, the audio-teacher model, and the visual-teacher model. The AVC student model, which consists of an acoustic student and a visual student, extracts the features from sound a_i and image v_j and uses a contrastive learning approach to learn a localization map o_{ij}^{xy}. This localization map identifies regions in the visual data that correlate with the sound signal. It is worth noting that the contrastive learning is based on a similarity computation that is performed for every possible audio-visual pair (a_i, v_j), where $1 \leq i \leq k$ and $1 \leq j \leq k$, and k is the number of instances in each sample batch. Simultaneously, while the AVC model learns from this contrastive approach, we utilize a knowledge distillation strategy to transfer object and acoustic knowledge to the AVC student model. Two teacher models, namely a sound [16] and an object recognition model [4] provide their learned features (feature-based hints) to supervise the AVC student. This supervision aims to ensure that the output features of the acoustic student and the visual student are respectively aligned with the features of existing sounds and objects in the input data. Intuitively, the feature-based hints of the teacher models provide valuable prior knowledge (i.e., class information of objects and sounds, as well as the spatial locations of objects within the scene) for audio-visual localization. The remaining sections explain the components of our proposed framework.

3.3 Audio-Visual Correspondence (AVC) model

The AVC model (Fig. 2) consists of two student models that processes the audio and visual inputs and aims to capture the correlations between audio signals and the corresponding visual cues. When audio is associated with a specific sound in the image, there will be visual cues in the image that correspond to that sound source. For example, if the sound source is a person speaking, visual cues like the presence of a person's face are likely correlated with the audio signal. Through training on a diverse dataset of audio-visual pairs, the model learns to identify and associate these visual cues with the corresponding audio features. By maximizing the agreement between the visual and audio representations of positive pairs, the model implicitly learns to identify the pixels in the image that emit the sound, as these pixels are responsible for the visual cues associated with the sound source.

The AVC model consists of two student models $F_a(\cdot)$ and $F_v(\cdot)$ that respectively process the sound a_i and image v_j, and obtain the corresponding f_{a_i} and f_{v_j} feature maps. Details of these encoders are presented in Section 3.6. We compute the cosine similarity between audio features f_{a_i} and visual features f_{v_j}, where $i, j = 1, \cdots, k$ and k is the number of samples in each sample batch. The dimensions of f_{a_i} and f_{v_j} are respectively $1 \times \ 1 \times \ c$ and $h \times \ w \times \ c$, in which h, w, and c stand for height, width, and channel dimensions of the feature maps. As a result, we obtain k^2 numbers of cross-modal feature maps (or output localization maps), each with a size of $h \times w$. Note that k of these feature maps (shown on the main diagonal of the $O_{k\times k}$ tensor in Fig. 2) are related to the same audio-image pair $\left\{(a_i, v_j)|i, j = 1, \ldots, n\right\}_{i=j}$, and the rest are related to different audio-image pairs $\left\{(a_i, v_j)|i, j = 1, \ldots, n\right\}_{i\neq j}$. We consider the maximum pixel value in each output map (i.e., each element of the $O_{k\times k}$ tensor) to build the cross-similarity matrix $S_{k\times k}$. We then use this matrix and its permuted version $S_{k\times k}^T$ to form a symmetric measure, and then we use a softmax activation followed by a cross-entropy loss function to optimize the similarity loss L_S as:

$$\mathcal{L}_S = -\log \frac{\exp(S_{k\times k}/\tau)}{\sum_k \exp(S_{k\times k}/\tau)} \cdot \frac{\exp(S_{k\times k}^T/\tau)}{\sum_k \exp(S_{k\times k}^T/\tau)}, \tag{1}$$

where τ is a hyperparameter acting as a smoothing factor [34]. This loss function was previously discussed e.g., in [8, 13, 19, 21] and is known as a contrastive learning loss function.

Note that, in the inference phase, we only compute the elements on the main diagonal of the $O_{k\times k}$ matrix, and extrapolate them with a bilinear operation to have the same resolution as the input image v_j.

3.4 Audio and visual teachers

The audio-visual localization task involves identifying the source of sounds that originate from visible objects within a scene. Therefore, excluding non-object sounds, such as those caused by wind or sounds that do not correspond to objects presented in the visual

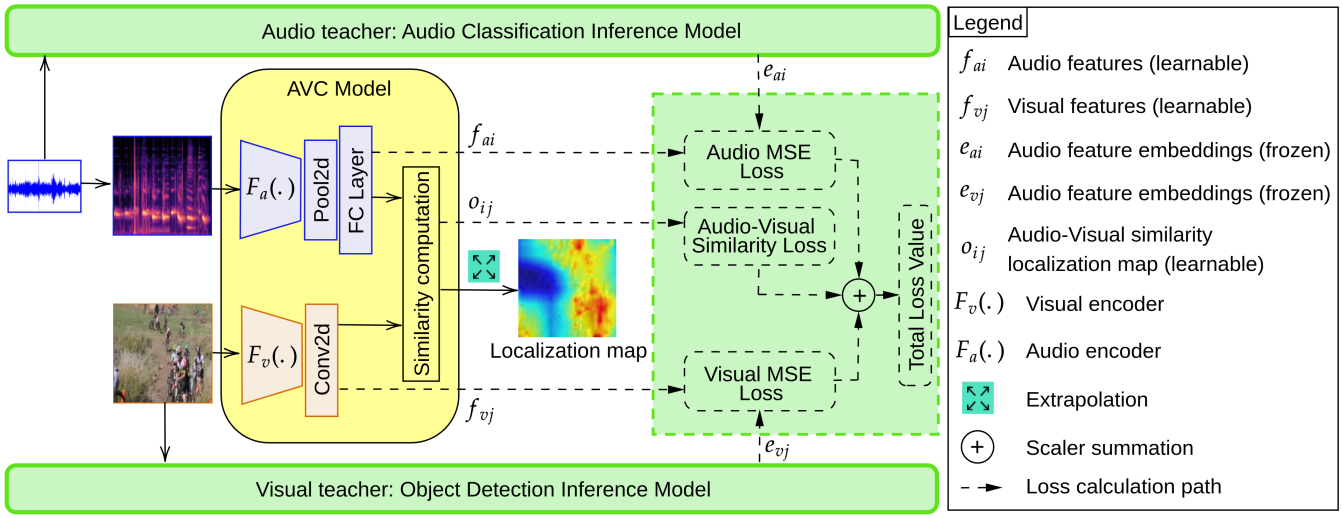

Figure 1: Overview of our proposed model. Two teachers (in green) supervise the training of the Audio-Visual Correspondence (AVC) model (in yellow), which generates a localization map as output. The green blocks are removed at the inference phase.

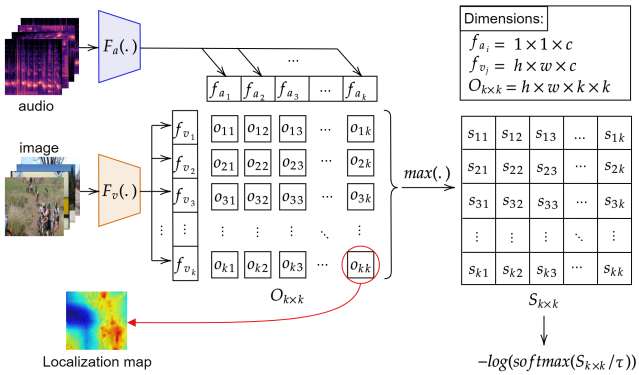

Figure 2: Details of the contrastive learning in the AVC model. Similarity is computed via the cosine similarity measure. The tensor $O_{k\times k}$ has four dimensions, so each of its elements (e.g., o_{kk} in the red circle) has two dimensions and provides a localization map. The maximum value of each localization map constructs a matrix $S_{k\times k}$, which is optimized using the cross entropy loss function.

data, can facilitate the learning of audio-visual localization. Having this in mind, we use an object detection model as a teaching model and transfer the object-based features to the visual student. While various object detectors are available, we opted for DETR [4] (DEtection TRansformer) in this study due to its publicly available pre-trained weights and user-friendly nature. DETR processes the input image v_j using a ResNet50 feature extractor, followed by an encoder-decoder transformer [37] that outputs the visual embedding e_{v_j}. We expect the embeddings e_{v_j} to include high-level semantic information about the scene, such as the class and location of the objects. Note that the knowledge distillation does not rely on object bounding boxes or class labels, instead, it uses the DETR feature maps to transfer these semantic knowledge. As in Fig. 1, we define a loss function to encourage the visual encoder $F_v(.)$ to learn from the existing semantic information in e_{v_j} by minimizing the mean squared error (squared L2 norm) between each element in f_{v_j} and our target embedding e_{v_j} as:

$$\mathcal{L}_{\text{v}} = \frac{1}{k}\sum_{j=1}^{k}(f_{v_j} - e_{v_j})^2, \tag{2}$$

where k is the number of samples in each sample batch. Following the same idea for the audio modality, we use the Pretrained Audio Neural Networks (PANNs) [16] as our audio teacher:

$$\mathcal{L}_{\text{a}} = \frac{1}{k}\sum_{i=1}^{k}(f_{a_i} - e_{a_i})^2. \tag{3}$$

The PANNs model is an audio pattern recognition framework that performs multi-label sound classification. This model combines the features extracted from raw waveforms and log-Mel spectrograms to obtain the audio embeddings e_{a_i}, while the audio student $F_a(.)$ extracts the features f_{a_i} based on a log-spectrogram input. log-Mel spectrogram is more condensed and is useful for capturing the overall characteristics, while a log-spectrogram focuses on fine details. Therefore, Eq. 3 helps the audio student to learn from overall characteristics of the audio signal.

We chose PANNs because it is trained on the Audio-set dataset [11] and has pre-trained weights with rich audio knowledge. Audio-set contains over 5000 hours of audio data with 527 sound classes, and samples have been collected from Youtube videos. Note that the datasets [3, 7] for the audio-visual localization task were collected from similar sources (i.e., Flickr and Youtube), assuring that our teacher and student models have access to a similar distribution of training data.

Overall, the AVC model is optimized using a loss function that combines the contrastive learning scheme (Eq. 1) with the visual and acoustic knowledge distillation approach (Eq. 2 and Eq. 3), and is defined as: $\mathcal{L}_{\text{total}} = \alpha_1\mathcal{L}_{\text{S}} + \alpha_2\mathcal{L}_{\text{a}} + \alpha_3\mathcal{L}_{\text{v}}$, where the hyper-parameters α_1 to α_3 control the relative impact of the contrastive learning component and each teacher model in the training process. To ensure that the contribution of each component

in the loss function is balanced, we set the hyper-parameter values to prevent the dominance of any term over the others. Specifically, we selected values of $\alpha_1 = 1$, $\alpha_2 = 1000$, and $\alpha_3 = 100{,}000$ to ensure that the loss terms fall within the same value range.

3.5 Object-guided inference

As suggested by EZ-VSL [21], an audio-visual localization model can benefit from an object-guided model in the inference phase. EZ-VSL [21] shows that a ResNet18 pre-trained on ImageNet improves the model performance in identifying the location of the acoustic signals in the visual data. Following this idea, we propose the object-guided variant of our AVC model (OG-AVC): Specifically, we use a standard ResNet18 model, excluding the last pooling and fully connected layers. We squeeze the features extracted from the ResNet backbone in the channel dimension using a mean operation and use a bilinear extrapolation to obtain an object-based localization map of the same size to the input image. To generate the output localization map, we use a weighted combination of two sources: the object-based localization map and the output of our AVC inference model. The weight for the object-based map is set to 0.4, while the AVC output is weighted at 0.6. Further details on this process are provided in the results section.

In an additional step, we integrated the predictions from our visual teacher, known as DETR [4], into our approach and proposed a variant of OG-AVC called OG-AVC-DETR. We considered the latest feature maps of DETR inference model that exhibited a confidence score. In particular, feature channels with confidence score of ≥ 0.3 indicate the presence of an object within the channel. Subsequently, we performed a mean operation on the remaining feature maps and compressed them. The resulting output was then extrapolated bilinearly, yielding a localization map with the same dimensions as the input image. By combining the predictions from the DETR model with ResNet18 and our proposed AVC model, each weighted equally at 1/3, we obtained the final output localization map. We compare the object-guided version of our model with EZ-VSL [21] in the next section.

3.6 Implementation details

Similar to recent works [6, 18, 21], we use the ResNet18 backbone in our AVC model to extract features from both modalities and follow the same data pre-processing procedure as in [6, 21]. It is worth mentioning that we used ViT [10] and different CNN encoders (ResNet, VGG [31], and EfficientNet [35]) on the 10k datasets and realized that our model works better when we use ResNet encoders; specifically, smaller models (ResNet18) could generalize better, probably because bigger models overfit to the training data.

For the student encoders $F_a(.)$ and $F_v(.)$, we employed the ResNet18 model with minor modifications in the final layers. Specifically, to design our $F_v(.)$, the average pooling and fully connected layer of ResNet18 are omitted, and instead, a 1×1 convolutional layer is introduced, transforming the 512-dimensional feature maps to 2048 dimensions. We input images of size 224×224 to $F_v(.)$ and obtain a feature map of size 7×7×2048.

To process audio signals, we use a mono-channel setup with a sampling rate of 22,050 Hz. Following [22] and [32], we focus on the central 3 seconds of each sound signal, which is equivalent to 66,150 sampling points. We selected 3-second chunks because it is probably sufficient to capture the main content of the audio that is temporally relevant to the visual frame while avoiding potentially irrelevant sounds. Next, similar to [6] and [21], we generate a log-spectrogram to enable the use of convolutional neural networks (CNN) for audio signal processing. We fed the log-spectrogram to the audio student model because it is simple to compute and contains the overall characteristics of the audio, which helps the audio student learn coarse audio features to be matched to the coarse visual cues obtained by the visual student. To generate the log-spectrogram, we segment the sampling points into windows of 512 points each, with adjacent windows overlapping by around 50% to reduce the occurrence of artifacts in the spectrogram. To visualize the spectrogram, we apply a logarithmic transformation to compress the amplitude values and obtain a log-spectrogram of dimension 1×257×276. Without applying any augmentations, we feed the log-spectrogram to the audio encoder $F_a(.)$ and obtain a feature map of size 1×1×512. The audio encoder $F_a(.)$ is a Resnet18 in which the first 2D convolution is modified to receive the single channel image (i.e., log-spectrogram of dimension 1×257×276) and the last average pooling and fully connected (FC) layers are removed, and instead, we add a max pooling layer followed by a FC layer with 2048 output channels. The reason to add this FC layer is to map the audio features to a shared feature space with the visual features. As a result of analyzing the input log-spectrogram of audio signals, we obtain a feature map of size 1×1×2048.

We map both modalities into a shared latent space using projection layers (Conv2d and FC layer shown in Fig. 1) with 2048 output channels. Then, we normalize the feature maps before applying Eq. 2 and Eq. 3. These equations require the pre-trained prior-knowledge embeddings e_{a_i} and e_{v_j} as well. To minimize training time and computation cost, we use the teacher models in an offline mode. Specifically, we extract the prior-knowledge embeddings by the teacher models and store them on the disk. During the training of the student model, we load these pre-extracted embeddings instead of running the teacher models to generate them repeatedly in each epoch.

We trained the model for 100 epochs with a batch size of 128, and considered the Adam optimizer with an initial learning rate of 0.0001, without a scheduler. The training was performed on a single NVIDIA GeForce RTX 4090 GPU with a power capacity of 450 Watt and a memory of 24,564 MebiBytes (MiB).

4 EVALUATION AND COMPARISON

4.1 Datasets and metrics

We trained our proposed framework on two 10k and 144k random subsets of Flicker [3] and VGGS [7] datasets. The Flickr dataset [3] consists of over 2 million unconstrained videos from Flickr with a length of 10 seconds, while the VGGS dataset [7] comprises over 200k video clips captured from Youtube. We then tested our model on the Flickr test set [28] and VGGSS set [6]. The Flickr test set [28] has 250 samples, while the VGGSS set [6] consists of 5k annotated samples. Both test sets have annotations for bounding boxes around the location of the emitted sounds. It is worth noting that we could evaluate the performance of our inference model based on 4600 samples of VGGSS [6], as the other instances were

Model	Train set	Test	cIoU	AUC
AVC-10k	Flickr-10k	Flickr	77.7±2.4	60.0±1.6
	Flickr-10k	VGGSS	29.4±1.7	35.0±0.7
	VGGS-10k	VGGSS	33.8±2.4	36.7±1.2
	VGGS-10k	Flickr	78.9±2.0	61.0±0.9
AVC-144k	Flickr-144k	Flickr	78.5±1.8	61.0±0.8
	Flickr-144k	VGGSS	31.5±3.4	35.8±1.4
	VGGS-144k	VGGSS	37.9±0.5	38.5±0.3
	VGGS-144k	Flickr	81.4±2.2	63.1±0.9

Table 1: The cross-dataset performance of the proposed AVC model. VGGSS [6] is the test set for VGGS [7]. 10k and 144k refer to the number of samples used for training the model. Values are based on 10 training and testing runs.

unavailable on Youtube. Over time, the number of available samples may even decrease. We evaluated our model on these datasets using a pre-processing approach described in [21, 28]. This involved extracting a 3-second audio signal and selecting one frame from the visual data at the 5-second mark. We utilized this audio-image pair to train our unsupervised learning method. Following [25, 28], we evaluate our model using the consensus Intersection over Union (cIoU) and Area Under Curve (AUC), in which a higher value indicates better performance, respectively, in terms of better localization and correctly identifying true positives while avoiding false positives. A detailed definition can be find at [28].

4.2 Evaluation and ablation studies

In this section, we evaluate our proposed model on the 10k and 144k subsets of both Flickr [3] and VGGS [7], including a cross-dataset evaluation, in which we interchangeably test the models on Flickr [28] and VGGSS [6]. Furthermore, we perform an ablation study to determine the impact of each component in the framework.

Table 1 shows the performance of the proposed AVC model on different train and test sets. The train sets are Flickr-10k, Flickr-144k, VGGS-10k, and VGGS-144k, and the test sets are either Flickr [28] or VGGSS [6]. Overall, the results suggest that the performance of the proposed AVC model is strongly influenced by the choice of training and test sets, as well as the size of the training set. For example, when trained on 144k samples, the model achieves better performance for all test sets compared to when trained on 10k samples. Additionally, Table 1 shows that the model performs significantly better on the Flickr test set by hovering around a cIoU of 80 and an AUC of 60, compared to the VGGSS test set, with cIoU and AUC scores of around 30 and 35, respectively. This suggests that the Flickr test set could be simpler than VGGSS. Another point is that the model achieves better results when trained on VGGS and tested on Flickr, regardless of the number of train samples, indicating that the model generalizes better when trained on VGGS sets rather than Flickr sets.

Table 2 shows the effects of ablating the audio and visual teachers separately and jointly in training phase of the AVC-10k model on the Flickr-10k dataset. Note that, as in knowledge distillation learning, the teacher models are only used for the training phase, and the inference model is the AVC model, i.e., the yellow block in Fig. 1. The first row of Table 2 shows the performance of the base model, which is the AVC model, trained only using a contrastive learning approach, without the supervision of any teacher. This suggests that

Mode	AVC	Audio teacher	Visual teacher	cIoU	AUC
Base	✓	-	-	67.4±3.8	56.3±1.7
AVC-10k	✓	✓	-	74.5±2.3	58.4±0.6
	✓	-	✓	71.6±5.2	57.9±2.3
	✓	✓	✓	77.7±2.4	59.9±1.5
OG-AVC-10k	✓	✓	✓	82.7±1.0	62.8±0.6

Table 2: Ablation studies in training phase: analyzing the impact of ablating the teachers on the AVC model. Teachers were removed at the inference phase. OG-AVC-10k is the object-guided variant of AVC-10k. The model was trained on the Flickr-10k and tested on the Flickr test set.

when the model only learns from the hidden correlation between audio and visual data to identify the location of the acoustic sound in the visual scene, the cIoU and AUC hover around 67.4 and 56.3, respectively; however, when the audio teacher helps the AVC student in the learning phase, the AVC student achieves 74.5 of the cIoU and 58.4 of the AUC. If we only use the visual teacher model to supervise the AVC student in the training phase, the inference AVC model obtains a cIoU of 71.6 and an AUC of 57.9, indicating that the visual teacher has a lower impact than the audio teacher on average. As expected, when both teachers supervise the learning process of the AVC student, the AVC model shows its highest performance with a cIoU of 77.7 and an AUC of 59.9.

The last row in Table 2, indicated as OG-AVC-10k, shows the results for an object-guided version of the AVC-10k model, in which a pixel-wise aggregation is performed between the prediction of the AVC student and that of an object-guided model, i.e., a ResNet18 pre-trained on ImageNet. The best results for OG-AVC-10k were obtained for a weighting factor of 0.4 for the output of the pre-trained ResNet18 model and a corresponding weighting of 0.6 for the output of the AVC inference model. As a result of this step, the model achieves a cIoU score of 82.7 and an AUC of 62.8.

Table 3 presents the detailed results of using the pre-trained ResNet18 in the inference phase, indicating that employing an object-guided model during the inference phase yields a significant improvement in the model performance (see Sec. 3.5). Specifically, comparing the results in Table 1 and Table 3 demonstrates that the object-guided version of AVC (OG-AVC) outperforms the original AVC model for all the experiments on both Flickr and VGGSS test sets, with an improvement of around 5 scores for cIoU and 3 scores for AUC. For example, AVC-10k trained and tested on Flickr achieves a cIoU of 77.7 and an AUC of 60.0, while OG-AVC-10k improves these scores to 82.7 and 62.8, respectively.

Our AVC model has 24.5 million parameters and performs about 7.2 billion floating-point operations (FLOPs) to process a single input. Employing ResNet18 as the object-guided model adds 11.2 million parameters and 1.8 billion FLOPs to the inference model.

4.3 Comparison with recent works

Table 4 shows the quantitative results of the comparison of our AVC model with regard to several existing methods [1, 6, 18, 23, 25, 28, 30, 32] based on two subsets of the Flickr dataset [3]: 10k and 144k. These works were selected as they are some of the most recent and relevant unsupervised approaches, proposing solutions to improve the contrasting learning strategy in audio-visual localization. When

Model	Train set	Test	cIoU	AUC
OG-AVC-10k	Flickr-10k	Flickr	82.7±1.0	62.8±0.6
	VGGS-10k	VGGSS	37.1±1.3	38.2±0.6
	Flickr-10k	VGGSS	34.5±1.1	37.2±0.4
	VGGS-10k	Flickr	82.0±1.2	62.7±0.4
OG-AVC-144k	Flickr-144k	Flickr	83.7±1.1	63.4±0.3
	VGGS-144k	VGGSS	39.2±0.5	39.2±0.3
	Flickr-144k	VGGSS	35.8±2.1	37.7±0.8
	VGGS-144k	Flickr	82.8±1.0	63.6±0.3

Table 3: The performance of the proposed model equipped with an object-guided model in the inference phase. OG-AVC indicates the object-guided version, in which ResNet18's prediction is aggregated with that of AVC. 10k and 144k refer to the number of samples used for training the model. VGGSS [6] is the test set for VGGS [7].

we train and test the proposed AVC model for 10 rounds on Flickr-10k, we achieve a cIoU of 77.7 and an AUC of 60.0, surpassing all the methods by at least a margin of 3.4 on cIoU and 1.3 on AUC. The evaluation results show that training on a larger set (Flickr-144k) improves the performance in respect of cIoU and AUC for all the methods; our proposed AVC-144k still surpasses all the methods [6, 14, 23, 28, 32] on cIoU by a margin of 2.6 and competes with [32] by achieving an equal AUC of 61.0.

Table 5 shows the comparison results of our AVC-10k and AVC-144k models with several recent works, when were trained on 10k and 144k samples of VGGS [7] and tested on VGGSS [6]. This table shows that our AVC-10k achieves a cIoU of 33.8 and an AUC of 36.7, which outperformed all other models [6, 28, 32] by a margin of 2.4 on cIoU and placed second best on AUC, after the SSPL [32] method with an AUC of 36.9. However, when the training data is increased to 144k samples, our model (AVC-144k) outperforms all the other methods [1, 6, 25, 28, 29, 32] by at least a margin of 3.3 on cIoU and 0.5 on AUC. The increase in cIoU scores by 3.3 demonstrates that our AVC-144k method has significantly improved the accuracy of audio-visual localization. Furthermore, the superior performance of our method, with a 0.5 AUC score, indicates its enhanced ability to effectively differentiate between positive and negative samples.

Train set	Model	cIoU	AUC
Flickr-10k	Attention [28]	43.6	44.9
	CoarsetoFine [25]	52.2	49.6
	AVObject [1]	54.6	50.4
	LVS [6]	58.2	52.5
	LM [23]	56.8	50.7
	USLICL [18]	71.0	58.0
	TDA [30]	73.4	57.6
	SSPL [32]	74.3	58.7
	AVC-10k† (ours)	**77.7**	**60.0**
Flickr-144k	Attention [28]	66.0	55.8
	DMC [14]	67.1	56.8
	LM [23]	68.4	57.0
	LVS [6]	69.9	57.3
	SSPL [32]	75.9	**61.0**
	AVC-144k† (ours)	**78.5**	**61.0**

Table 4: Comparison results with several state-of-the-art models on the Flickr dataset. The models were trained on random subsets (10k and 144k) of the Flickr dataset [3] and tested on a test set with 250 annotated samples [6]. † shows the average results obtained from 10 train and test runs of our AVC model. The best results are bold.

Train set	Model	cIoU	AUC
VGGS-10k	Attention [28]	16.0	28.3
	LVS [6]	27.7	34.9
	SSPL [32]	31.4	**36.9**
	AVC-10k† (ours)	**33.8**	36.7
VGGS-144k	Attention [28]	18.5	30.2
	CoarsetoFine [25]	29.1	34.8
	AVObject [1]	29.7	35.7
	SSPL [32]	33.9	38.0
	LVS [6]	34.4	38.2
	HardPos [29]	34.6	38.0
	AVC-144k† (ours)	**37.9**	**38.5**

Table 5: Comparison results on the VGGS dataset. The models were trained on random subsets (10k and 144k) of VGGS [7] and tested on VGGSS [6]. † shows the average results of 10 train and test runs. The best results are bold.

4.4 Cross-dataset evaluation

Table 6 shows the comparison results between our proposed model and recent methods [6, 28, 32], when all the methods are trained on VGGS-10k and VGGS-144k and tested on Flickr test set [28]. This cross-dataset validation shows that our model outperforms all other models, with a cIoU of 78.9 and an AUC of 61.0, when trained on VGGS-10k, indicating that our model surpasses the second-best method (SSPL [32]) by a margin of 2.6 and 1.1, respectively. On VGGS-114k, the performance improvement of our method over SSPL [32] increases to 3.7 for cIoU and 2.6 for AUC.

4.5 Object-guided variant evaluation

As discussed in section 3.5, the audio-visual localization models can benefit from an object-guided model in the inference phase, which is suggested by EZ-VSL [21]. To have a fair comparison with EZ-VSL [21], we used the same settings and object-guided model (i.e., a pre-trained ResNet18) in the inference phase. Table 7 shows that our approach improves the cIoU from 81.9 to 82.7 and the AUC from 62.6 to 62.8, when we train both methods on Flickr-10k.

Train set	Test set	Model	cIoU	AUC
VGGS-10k	Flickr	Attention* [28]	52.2	50.2
		LVS [6]	61.8	53.6
		SSPL [32]	76.3	59.1
		AVC† (ours)	**78.9**	**61.0**
VGGS-144k	Flickr	LVS [6]	71.9	58.2
		SSPL [32]	76.7	60.5
		AVC† (ours)	**81.4**	**63.1**

Table 6: Performance for cross-dataset validation. † shows the average results of 10 train and test runs. * denotes that the results were taken from [32]. The best results are bold.

Train set	Test	Model	cIoU	AUC
Flickr-10k	Flickr	EZ-VSL [21]	81.9	62.6
	Flickr	OG-AVC-10k †	**82.7**	**62.8**
Flickr-144k	Flickr	EZ-VSL [21]	83.1	63.1
	Flickr	OG-AVC-144k †	**83.7**	**63.4**
VGGS-144k	VGGSS	EZ-VSL [21]	38.9	**39.5**
	VGGSS	OG-AVC-144k †	**39.2**	39.2

Table 7: Performance when an object-guided model (i.e., a pre-trained ResNet18 on ImageNet) cooperates with the inference model. † shows the average results of our object-guided version of AVC in the inference phase for 10 runs. The best results are bold.

Our OG-AVC-144k achieves a cIoU of 83.7 and an AUC of 63.4, while EZ-VSL [21] falls behind our method, with a cIoU of 83.1 and an AUC of 63.1. Moreover, our model is also evaluated on the VGGS-144k dataset, where it achieves better results than EZ-VSL [21] in terms of cIoU, while the AUC decreases by 0.3 scores.

We tested our object-guided model with different pre-trained models of ResNet (ResNet18, ResNet34, ResNet50, ResNet101), VGG16, and EfficientNet, and achieved the best results when a ResNet18 is considered as the object-guided model. In a further step, we also incorporated the predictions of the visual teacher model (i.e., DETR [4]) in the inference phase. This means that the final localization map is obtained from a weighted addition (with a weighting factor of 1/3) of three models. As a result of this experiment, the quantitative results of the OG-AVC remained almost the same for Flickr test set [28]; however, the qualitative results showed a meaningful improvement (see Fig. 3). As the DETR model has 36.8 million parameters and requires 57.1 billion FLOPs, our OG-AVC-DETR variant has 61.3 million parameters and 64.3 billion FLOPs in the inference phase.

4.6 Qualitative results

Figure 3 shows the qualitative performance of our AVC-144k model, the object-guided variant (OG-AVC-144k), and its extension based on the DETR object detection model [4] (i.e., OG-AVC-DETR) on a few samples from the Flickr test set [6]. Note that the input to the models is one image along with a 3-second audio signal. The ground truth annotation is a bounding box around the location of the sound in the image, and each image has been annotated by 3 annotators, while each annotator may annotate multiple sound locations. The qualitative results indicate that the AVC-144k model could predict the sound locations in the image successfully. When we incorporate the pre-trained ResNet18 in the inference phase, our model (OG-AVC-144k) shows a slight improvement in identifying the sound location in the first two rows in Fig. 3, while it misleads the model in the examples shown in the last two rows. The last column shows the prediction of the proposed method, in which we aggregate the predictions of a pre-trained ResNet18 and an object detection model (DETR [4]) with our AVC-144k student model, each with a weight of 1/3. The qualitative results show that the object detection model (DETR [4]) can help to predict smaller and more precise locations of sounds in the image; however, there are samples, such as the last row in Fig. 3, in which the object detection model misleads the inference model. These types of examples are much more frequent in VGGSS [6], making the use of a precise object detector ineffective for audio-visual localization on VGGSS.

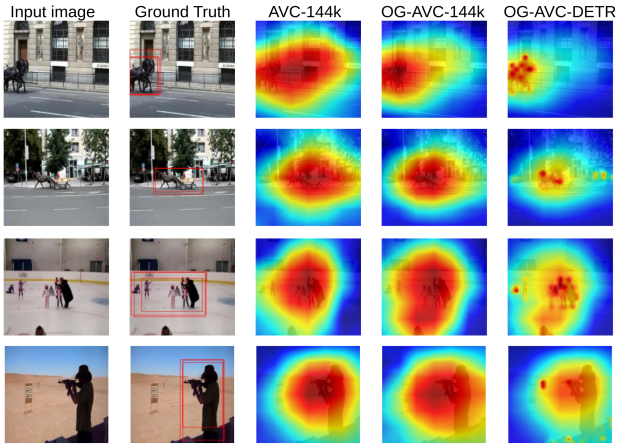

Figure 3: Localization maps for our variants: AVC-144k (student model), OG-AVC (object-guided AVC-144k), and OG-AVC-DETR (AVC-144k with ResNet18 and DETR predictions). Trained on Flickr-144k, samples from test set.

Overall, our qualitative results indicate that the proposed AVC model successfully predicts the sound location in the visual data, and incorporating the pre-trained ResNet18 into the inference model causes an incremental improvement. Further, the integration of an object detection model (DETR [4]) enables more precise localization in most cases, especially in the Flickr test set.

5 CONCLUSION AND OUTLOOK

In this paper, we present an unsupervised student-teacher architecture for audio-visual localization. Our main idea is to transfer information from pre-trained object detection and sound classification models to a vanilla audio-visual correspondence model and learn from existing knowledge without the need for labeled data. Our qualitative and quantitative experiments on two subsets of the VGGS and Flickr datasets showed that the proposed method outperforms the current state of the art on both datasets. The results indicate that learning from both sound and object prior knowledge is able to significantly improve audio-visual localization. Additional experiments suggested that incorporating the predictions of a pre-trained ResNet18 and an object detection method (DETR) in the inference phase can further improve performance. For future research, the knowledge transfer could be improved, e.g., by adding hints from the intermediate and last feature maps of the teachers [15] and by using attention mechanisms that indirectly transfer the useful knowledge [38].

ACKNOWLEDGMENTS

This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) with Project-ID 261402652 and the National Science Foundation of China with Project-ID 62061136001 in project Crossmodal Learning, TRR-169.

REFERENCES

[1] Triantafyllos Afouras, Andrew Owens, Joon Son Chung, and Andrew Zisserman. 2020. Self-supervised learning of audio-visual objects from video. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*. Springer, Cham, 208–224.

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

[3] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. 2016. SoundNet: Learning Sound Representations from Unlabeled Video. In *Advances in Neural Information Processing Systems* (Barcelona, Spain) *(NIPS'16)*. Curran Associates Inc., Red Hook, NY, USA, 892–900.

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *Computer Vision–ECCV 2020: 16th European Conference, August 23–28, 2020, Proceedings, Part I 16*. Springer, Glasgow, UK, 213–229.

[5] M.C. Casey, A. Pavlou, and A. Timotheou. 2012. Audio-visual localization with hierarchical topographic maps: Modeling the superior colliculus. *Neurocomputing* 97 (2012), 344–356. https://doi.org/10.1016/j.neucom.2012.05.015

[6] 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*. IEEE, New York, USA, 16867–16876.

[7] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. 2020. Vggsound: A Large-Scale Audio-Visual Dataset. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, New York, USA, 721–725. https://doi.org/10.1109/ICASSP40776.2020.9053174

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PMLR, Vienna, Austria, 1597–1607.

[9] Yanbei Chen, Yongqin Xian, A. Sophia Koepke, Ying Shan, and Zeynep Akata. 2021. Distilling Audio-Visual Knowledge by Compositional Contrastive Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, New York, USA, 7016–7025.

[10] 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).

[11] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, New York, USA, 776–780.

[12] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. 2021. Knowledge distillation: A survey. *International Journal of Computer Vision* 129 (2021), 1789–1819.

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

[14] 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*. IEEE, New York, USA, 9248–9257.

[15] Xiao Jin, Baoyun Peng, Yichao Wu, Yu Liu, Jiaheng Liu, Ding Liang, Junjie Yan, and Xiaolin Hu. 2019. Knowledge distillation via route constrained optimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. IEEE, New York, USA, 1345–1354.

[16] 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.

[17] L Ashok Kumar, D Karthika Renuka, S Lovelyn Rose, M C Shunmuga priya, and I Made Wartana. 2022. Deep learning based assistive technology on audio visual speech recognition for hearing impaired. *International Journal of Cognitive Computing in Engineering* 3 (2022), 24–30. https://doi.org/10.1016/j.ijcce.2022.01.003

[18] Yan-Bo Lin, Hung-Yu Tseng, Hsin-Ying Lee, Yen-Yu Lin, and Ming-Hsuan Yang. 2023. Unsupervised sound localization via iterative contrastive learning. *Computer Vision and Image Understanding* 227 (2023), 103602.

[19] Ishan Misra and Laurens van der Maaten. 2020. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. IEEE, New York, USA, 6707–6717.

[20] 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., New Orleans, Louisiana, USA, 37524–37536. https://proceedings.neurips.cc/paper_files/paper/2022/file/f3f2ff9579ba6deeb89caa2fe1f0b99c-Paper-Conference.pdf

[21] Shentong Mo and Pedro Morgado. 2022. Localizing Visual Sounds the Easy Way. In *Computer Vision – ECCV 2022*, Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (Eds.). Springer Nature Switzerland, Cham, 218–234.

[22] Quan Nguyen, Julius Richter, Mikko Lauri, Timo Gerkmann, and Simone Frintrop. 2021. Improving mix-and-separate training in audio-visual sound source separation with an object prior. In *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, New York, USA, 5844–5851.

[23] Takashi Oya, Shohei Iwase, Ryota Natsume, Takahiro Itazuri, Shugo Yamaguchi, and Shigeo Morishima. 2020. Do we need sound for sound source localization?. In *Proceedings of the Asian Conference on Computer Vision* (Kyoto, Japan). Springer, Cham, 119–136.

[24] Andres Perez, Valentina Sanguineti, Pietro Morerio, and Vittorio Murino. 2020. Audio-visual model distillation using acoustic images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. IEEE, New York, USA, 2854–2863.

[25] 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*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.). Springer International Publishing, Cham, 292–308.

[26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, Virtual, 8748–8763.

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

[28] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. 2021. Learning to Localize Sound Sources in Visual Scenes: Analysis and Applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, 5 (2021), 1605–1619. https://doi.org/10.1109/TPAMI.2019.2952095

[29] Arda Senocak, Hyeonggon Ryu, Junsik Kim, and In So Kweon. 2022. Learning Sound Localization Better from Semantically Similar Samples. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, New York, USA, 4863–4867. https://doi.org/10.1109/ICASSP43922.2022.9747867

[30] 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*. IEEE, New York, USA, 1737–1746.

[31] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[32] 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*. IEEE, New York, USA, 3222–3231.

[33] Jacek Stachurski, Lorin Netsch, and Randy Cole. 2013. Sound source localization for video surveillance camera. In *2013 10th IEEE International Conference on Advanced Video and Signal Based Surveillance*. IEEE, New York, USA, 93–98. https://doi.org/10.1109/avss.2013.6636622

[34] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. IEEE, New York, USA, 2818–2826.

[35] Mingxing Tan and Quoc Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*. PMLR, 6105–6114.

[36] 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*. IEEE, New York, USA, 2745–2754.

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc., Long Beach, CA, USA. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[38] Sergey Zagoruyko and Nikos Komodakis. 2017. Paying More Attention to Attention: Improving the Performance of Convolutional Neural Networks via Attention Transfer. In *International Conference on Learning Representations*. HAL open science, Paris, France, 1–13. https://openreview.net/forum?id=Sks9_ajex

[39] Xinchi Zhou, Dongzhan Zhou, Di Hu, Hang Zhou, and Wanli Ouyang. 2023. Exploiting Visual Context Semantics for Sound Source Localization. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, New York, USA, 5188–5197. https://doi.org/10.1109/WACV56688.2023.00517

[40] Hao Zhu, Man-Di Luo, Rui Wang, Ai-Hua Zheng, and Ran He. 2021. Deep audio-visual learning: A survey. *International Journal of Automation and Computing* 18 (2021), 351–376.