Audio-Visual Grouping Network for Sound Localization from Mixtures

Shentong Mo
Carnegie Mellon University

Yapeng Tian*
University of Texas at Dallas

Abstract

Sound source localization is a typical and challenging task that predicts the location of sound sources in a video. Previous single-source methods mainly used the audio-visual association as clues to localize sounding objects in each image. Due to the mixed property of multiple sound sources in the original space, there exist rare multi-source approaches to localizing multiple sources simultaneously, except for one recent work using a contrastive random walk in the graph with images and separated sound as nodes. Despite their promising performance, they can only handle a fixed number of sources, and they cannot learn compact class-aware representations for individual sources. To alleviate this shortcoming, in this paper, we propose a novel audio-visual grouping network, namely AVGN, that can directly learn category-wise semantic features for each source from the input audio mixture and image to localize multiple sources simultaneously. Specifically, our AVGN leverages learnable audio-visual class tokens to aggregate class-aware source features. Then, the aggregated semantic features for each source can be used as guidance to localize the corresponding visual regions. Compared to existing multi-source methods, our new framework can localize a flexible number of sources and disentangle category-aware audio-visual representations for individual sound sources. We conduct extensive experiments on MUSIC, VGGSound-Instruments, and VGG-Sound Sources benchmarks. The results demonstrate that the proposed AVGN can achieve state-of-the-art sounding object localization performance on both single-source and multi-source scenarios. Code is available at https://github.com/stoneMo/AVGN.

1. Introduction

When we hear a dog barking, we are naturally aware of where the dog is in the room due to the strong correspondence between audio signals and visual objects in the world. In the meanwhile, we are capable of separating individual sources from a mixture of multiple sources in the daily envi-
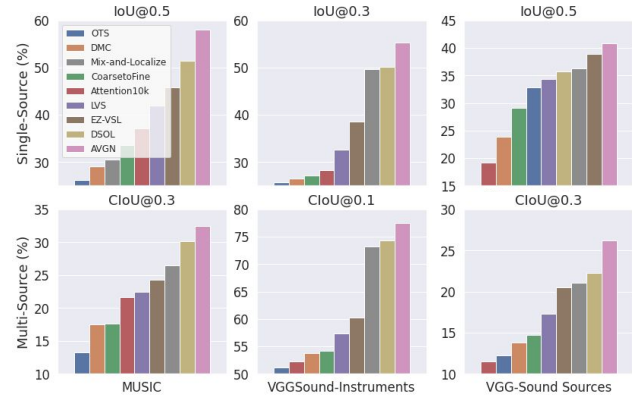
---
*Corresponding author.

Figure 1. Comparison of our AVGN with state-of-the-art methods on single-source (Top Row) and multi-source (Bottom Row) sound localization on MUSIC [51], VGGSound-Instruments [23], and VGG-Sound Sources [9] benchmarks.

ronment. This human perception intelligence attracts many researchers to explore audio-visual joint learning for visual sound source localization.

Visual sound source localization is a typical and challenging task that predicts the location of sound sources in a video. To tackle this problem, early single-source methods [1,9,21,30,31,39,40] mainly used the audio-visual association as clues to localize sounding objects in the frame. Typically, Attention10k [39] introduced a two-stream architecture for audio and images to localize a sound source in the image using an attention mechanism. Based on the attention, Afouras *et al.* [1] incorporated the optical flow for more accurate localization in a video. To explicitly learn discriminative audio-visual corresponding fragments, LVS [9] proposed hard sample mining with a differentiable threshold-based contrastive loss, while HardPos [40] utilized hard positives from negative pairs in the loss. More recently, EZVSL [31] developed a multiple-instance contrastive learning framework on the most aligned regions corresponding to the audio. SLAVC [30] adopted momentum encoders and extreme visual dropout to address overfitting and silence issues in single-source sound localization. However, those baselines are based on the single-source sound as input and they achieve worse performance for sound localization from mixtures. In this work, we will solve the problem in our approach by extracting disentangled and compact

representations with learnable audio-visual class tokens as guidance for sound localization.

Since multiple sound sources are mixed in the original space, recent researchers have tried to explore diverse pipelines to localize multiple sources on frames from a sound mixture. This multi-source task requires the model to associate individual sources separated from the mixture with each frame. Qian *et al.* [38] leveraged a two-stage framework to capture cross-modal feature alignment between sound and vision representations in a coarse-to-fine manner. DSOL [22] introduced a two-stage training framework to tackle with silence in category-aware sound source localization. More recently, Mix-and-Localize [23] proposed to use a contrastive random walk in the graph with images and separated sounds as nodes, where a random walker was trained to walk from each audio node to an image node with audio-visual similarity as the transition probability. Despite their promising performance, they can only handle a fixed number of sources and they cannot learn compact class-aware representations for individual sources. In contrast, we can support a flexible number of sources as input and learn class-aware representations for each source.

The main challenge is that sounds are naturally mixed in the audio space. This inspires us to disentangle the individual semantics for each source from the mixture to guide source localization. To address the problem, our key idea is to disentangle individual source representation using audio-visual grouping for source localization, which is different from existing single-source and multi-source methods. During training, we aim to learn audio-visual category tokens to aggregate category-aware source features from the sound mixture and the image, where separated high-level semantics for individual sources are learned.

To this end, we propose a novel audio-visual grouping network, namely AVGN, that can directly learn category-wise semantic features for each source from the input audio mixture and frame to localize multiple sources simultaneously. Specifically, our AVGN leverages learnable audio-visual class tokens to aggregate class-aware source features. Then, the aggregated semantic features for each source will serve as guidance to localize the corresponding visual regions. Compared to previous multi-source baselines, our new framework can support a flexible number of sources and shows the effectiveness of learning compact audio-visual representations with category-aware semantics.

Empirical experiments on MUSIC and VGGSound-Instruments benchmarks comprehensively demonstrate the state-of-the-art performance against previous single-source and multi-source baselines. In addition, qualitative visualizations of localization results vividly showcase the effectiveness of our AVGN in localizing individual sources from mixtures. Extensive ablation studies also validate the importance of category-aware grouping and learnable audio-visual class tokens in learning compact representations for sound source localization.

Our main contributions can be summarized as follows:

- We present a novel Audio-Visual Grouping Network, namely AVGN, to disentangle the individual semantics from sound mixtures and images to guide source localization.

- We introduce learnable audio-visual class tokens and category-aware grouping in sound localization to aggregate category-wise source features with explicit high-level semantics.

- Extensive experiments comprehensively demonstrate the state-of-the-art superiority of our AVGN over previous baselines on both single-source and multi-source sounding object localization.

2. Related Work

Audio-Visual Joint Learning. Audio-visual joint learning has been addressed in many previous works [2, 4, 12, 14, 19, 21, 26, 33, 34, 36, 37, 39, 50, 51] to learn the audio-visual correlation between two distinct modalities from videos. Such cross-modal alignments are beneficial for many audio-visual tasks, such as audio-event localization [27, 29, 43, 47], audio-visual spatialization [6, 17, 33, 35], audio-visual navigation [6–8] and audio-visual parsing [28, 32, 42, 46]. In this work, our main focus is to learn compact audio-visual representations for localizing individual sources on images from sound mixtures, which is more demanding than the tasks aforementioned above.

Audio-Visual Source Separation. Audio-visual source separation aims to separate individual sound sources from the audio mixture given the image with sources on it. In recent years, researchers [12, 16, 19, 41, 44, 49, 51] have tried to explore diverse pipelines to learn discriminative visual representations from images for source separation. Zhao *et al.* [51] first proposed a "Mix-and-Separate" network to capture the alignment between pixels and the spectral components of audio for the reconstruction of each input source spectrogram. With the benefit of visual cues, MP-Net [49] utilized a recursive MinusPlus Net to separate all salient sounds from the mixture. Tian *et al.* [41] used a cyclic co-learning framework with sounding object visual grounding to separate visual sound sources. More recently, more types of modalities have been explored to boost the performance of audio-visual source separation, such as motion in SoM [50], gesture composed of pose and keypoints in MG [15], and spatio-temporal visual scene graphs in AVSGS [5]. Different from them, we do not need to recover the audio spectrogram of individual sources from the mixture. Instead, we leverage the category-aware representations of individual sources to localize the corresponding
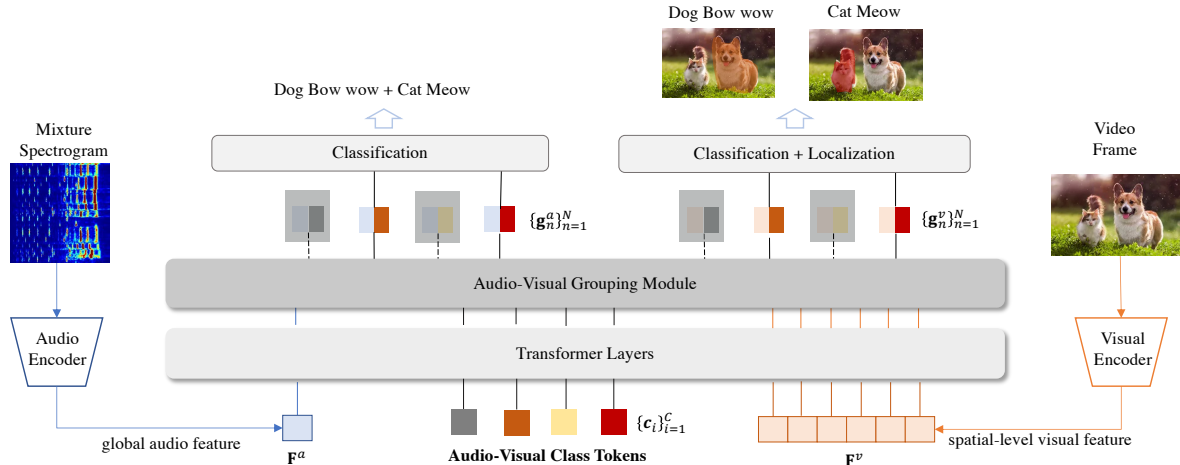
Figure 2. Illustration of the proposed Audio-Visual Grouping Network (AVGN). The Audio-Visual Grouping module takes as global audio $\mathbf{F}^a = \mathbf{f}^a$ of the mixture spectrogram, spatial-level visual features $\mathbf{F}^v = \{\mathbf{f}_p^v\}_{p=1}^P$ of the video frame from each encoder and learnable audio-visual class tokens $\{\mathbf{c}_i\}_{i=1}^C$ of for C categories in the semantic space to generate disentangled class-aware audio-visual representations $\{\mathbf{g}_n^a\}_{n=1}^N, \{\mathbf{g}_n^v\}_{n=1}^N$ for N sources. Note that N source embeddings are chosen from C categories according to the ground-truth class. Finally, two classification layers composed of an FC layer and a sigmoid function are separately used to predict audio and video categories, and localization maps are generated by the cosine similarity between audio-visual class-aware embeddings.

regions for each source, where learnable audio-visual class tokens are applied as the desirable guidance.

Visual Sound Source Localization. Visual sound source localization is a typical and challenging problem that predicts the location of individual sound sources in a video. Early works [13, 20, 24] applied traditional machine learning approaches, such as statistical models [13] and canonical correlation analysis [24] to learn low-level alignment between audio and visual representations. With the success of deep neural nets, recent researchers [1, 9, 21, 30, 31, 38–40] explored many architectures to learn the audio-visual correspondence for localizing single-source sounds. Attention10k [39] localized a sound source in the image using a two-stream architecture with an attention mechanism. Hard sample mining was introduced in LVS [9] to optimize a differentiable threshold-based contrastive loss for predicting discriminative audio-visual correspondence maps. More recently, a multiple-instance contrastive learning framework was proposed in EZVSL [31] to align regions with the most corresponding audio without negative regions involved.

Due to the natural mixed property of sounds in our environment, recent works [22, 23, 38] also have explored different frameworks to localize multiple sources on frames from a sound mixture simultaneously. DSOL [22] utilized a two-stage training framework to deal with silence for category-aware sound source localization. More recently, a contrastive random walk model was trained in Mix-and-Localize [23] to link each audio node with an image node using a transition probability of audio-visual similarity. While those single-source and multi-source approaches achieve promising performance in sound local-

ization, they can only handle a fixed number of sources and they cannot learn discriminative class-aware representations for individual sources. In contrast, we develop a fully novel framework to aggregate compact category-wise audio and visual source representations with explicit learnable source class tokens. To the best of our knowledge, we are the first to leverage an explicit grouping mechanism for sound source localization. Our experiments in Section 4.2 also demonstrate the effectiveness of AVGN in both single-source and multi-source localization.

3. Method

Given an image and a mixture of audio, our target is to localize individual sound sources on the image. We propose a novel Audio-Visual Grouping Network named AVGN for disentangling individual semantics from the mixture and image, which mainly consists of two modules, Audio-Visual Class Tokens in Section 3.2 and Audio-Visual Grouping in Section 3.3.

3.1. Preliminaries

In this section, we first describe the problem setup and notations, and then revisit the multiple-instance contrastive learning in EZVSL [31] for single-source localization.

Problem Setup and Notations. Given a mixture spectrogram and an image, our goal is to localize N individual sound sources in the image spatially. For a video with C source event categories, we have an audio-visual label, which is denoted as $\{y_i\}_{i=1}^C$ with y_i for the ground-truth category entry i as 1. During the training, we do not have bounding boxes and mask-level annotations. Therefore, we can only use the video-level label for the mixture spectrogram and image to perform weakly-supervised learning.

Revisit Single-source Localization. To address the single-source localization problem, EZ-VSL [31] introduced a multiple-instance contrastive learning framework to align the audio and visual features at locations corresponding to sound sources. Given a set of global audio feature $\mathbf{F}^a = \mathbf{f}^a \in \mathbb{R}^{1 \times D}$ and spatial-level visual features spanning all locations in an image $\mathbf{F}^v = \{\mathbf{f}^v_p\}^P_{p=1}, \mathbf{f}^v_p \in \mathbb{R}^{1 \times D}$, EZ-VSL applied the multiple-instance contrastive objective to align at least one location in the corresponding bag of visual features with the audio representation in the same mini-batch, which is defined as:

$$\mathcal{L}_{\text{baseline}} = -\frac{1}{B} \sum_{b=1}^{B} \log \frac{\exp\left(\frac{1}{\tau}\text{sim}(\mathbf{F}^a_b, \mathbf{F}^v_b)\right)}{\sum_{m=1}^{B} \exp\left(\frac{1}{\tau}\text{sim}(\mathbf{F}^a_b, \mathbf{F}^v_m)\right)} \tag{1}$$

where the similarity $\text{sim}(\mathbf{F}^a, \mathbf{F}^v)$ denotes the max-pooled audio-visual cosine similarity of $\mathbf{F}^a$ and $\mathbf{F}^v = \{\mathbf{f}^v_p\}^P_{p=1}$ across all P spatial locations. B is the batch size, D is the dimension size, and τ is a temperature hyper-parameter.

However, such a training objective will pose the main challenge for multi-source localization. The global audio representation extracted from the mixture is mixed and thus they can not associate individual sources separated from the mixture with the corresponding regions. To address the challenge, we are inspired by [48] and propose a novel Audio-Visual Grouping Network that can learn to disentangle the individual semantics from the mixture and image to guide multi-source localization, as illustrated in Figure 2.

3.2. Audio-Visual Class Tokens

In order to explicitly disentangle individual semantics from the mixed sound space and image, we introduce a novel learnable audio-visual class tokens $\{\mathbf{c}_i\}^C_{i=1}$ to help group semantic-aware information from audio-visual representations $\mathbf{f}^a, \{\mathbf{f}^v_p\}^P_{p=1}$, where $\mathbf{c}_i \in \mathbb{R}^{1 \times D}$, C is the total number of source classes, P denotes the number of total locations in the spatial map.

With the categorical audio-visual tokens and raw representations, we first apply self-attention transformers $\phi^a(\cdot), \phi^v(\cdot)$ to aggregate global audio and spatial visual features from the raw input and align the features with the categorical token embeddings as:

$$\hat{\mathbf{f}}^a, \{\hat{\mathbf{c}}^a_i\}^C_{i=1} = \{\phi^a(\mathbf{x}^a_j, \mathbf{X}^a, \mathbf{X}^a)\}^{1+C}_{j=1}, \\ \mathbf{X}^a = \{\mathbf{x}^a_j\}^{1+C}_{j=1} = [\mathbf{f}^a; \{\mathbf{c}_i\}^C_{i=1}] \tag{2}$$

$$\{\hat{\mathbf{f}}^v_p\}^P_{p=1}, \{\hat{\mathbf{c}}^v_i\}^C_{i=1} = \{\phi^v(\mathbf{x}^v_j, \mathbf{X}^v, \mathbf{X}^v)\}^{P+C}_{j=1}, \\ \mathbf{X}^v = \{\mathbf{x}^v_j\}^{P+C}_{j=1} = [\{\mathbf{f}^v_p\}^P_{p=1}; \{\mathbf{c}_i\}^C_{i=1}] \tag{3}$$

where $[\ ;\ ]$ denotes the concatenation operator. $\hat{\mathbf{f}}^a, \hat{\mathbf{f}}^v_p, \hat{\mathbf{c}}^a_i, \hat{\mathbf{c}}^v_i \in \mathbb{R}^{1 \times D}$, and and D is the dimension of embeddings. The self-attention operators $\phi^a(\cdot)$ is

formulated as:

$$\phi^a(\mathbf{x}^a_j, \mathbf{X}^a, \mathbf{X}^a) = \text{Softmax}(\frac{\mathbf{x}^a_j \mathbf{X}^{a\top}}{\sqrt{D}})\mathbf{X}^a \tag{4}$$

$$\phi^v(\mathbf{x}^v_j, \mathbf{X}^v, \mathbf{X}^v) = \text{Softmax}(\frac{\mathbf{x}^v_j \mathbf{X}^{v\top}}{\sqrt{D}})\mathbf{X}^v \tag{5}$$

Then, in order to constrain the independence of each class token $\mathbf{c}_i$ in the audio-visual semantic space, we apply a fully-connected (FC) layer and add softmax operator to predict the individual source class probability: $\mathbf{e}_i = \text{Softmax}(\text{FC}(\mathbf{c}_i))$. Each audio-visual category probability is optimized by a cross-entropy loss $\sum_{i=1}^{C} \text{CE}(\mathbf{h}_i, \mathbf{e}_i)$, where $\text{CE}(\cdot)$ is cross-entropy loss; $\mathbf{h}_i$ denotes a one-hot encoding with its element for the target category entry i as 1. Optimizing the loss will push the learned token embeddings to be category-aware and discriminative.

3.3. Audio-Visual Grouping

With the benefit of the aforementioned category-constraint objective, we propose a novel and explicit audio-visual grouping module composed of grouping blocks $g^a(\cdot), g^v(\cdot)$ to take the learned audio-visual source class tokens and aggregated features as inputs to generate category-aware audio-visual embeddings as:

$$\{\mathbf{g}^a_i\}^C_{i=1} = g^a(\{\hat{\mathbf{f}}^a, \{\hat{\mathbf{c}}^a_i\}^C_{i=1}), \\ \{\mathbf{g}^v_i\}^C_{i=1} = g^v(\{\hat{\mathbf{f}}^v_p\}^P_{p=1}, \{\hat{\mathbf{c}}^v_i\}^C_{i=1}) \tag{6}$$

During the phase of grouping, we merge all the audio-visual features from the same class token into a new class-aware audio-visual feature, by calculating the global audio similarity vector $\mathbf{A}^a \in \mathbb{R}^{1 \times C}$ and spatial visual similarity matrix $\mathbf{A}^v \in \mathbb{R}^{P \times C}$ between audio-visual features and audio-visual class tokens via a softmax operation, which is formulated as

$$\mathbf{A}^a_i = \text{Softmax}(W^a_q \hat{\mathbf{f}}^a \cdot W^a_k \hat{\mathbf{c}}^a_i), \\ \mathbf{A}^v_{p,i} = \text{Softmax}(W^v_q \hat{\mathbf{f}}^v_p \cdot W^v_k \hat{\mathbf{c}}^v_i) \tag{7}$$

where $W^a_q, W^a_k \in \mathbb{R}^{D \times D}$ and $W^v_q, W^v_k \in \mathbb{R}^{D \times D}$ denote the learnable weights of linear projections for the features and class tokens of audio and visual modalities, separately. With this global audio similarity vector and spatial visual similarity matrix, we compute the weighted sum of all global audio and spatial visual features assigned to generate the category-aware representations as:

$$\mathbf{g}^a_i = g^a(\hat{\mathbf{f}}^a, \hat{\mathbf{c}}^a_i) = \hat{\mathbf{c}}^a_i + W^a_o \frac{\mathbf{A}^a_i W^a_v \hat{\mathbf{f}}^a}{\mathbf{A}^a_i}, \\ \mathbf{g}^v_i = g^v(\{\hat{\mathbf{f}}^v_p\}^P_{p=1}, \hat{\mathbf{c}}^v_i) = \hat{\mathbf{c}}^v_i + W^v_o \frac{\sum_{p=1}^{P} \mathbf{A}^v_{p,i} W^v_v \hat{\mathbf{f}}^v_p}{\sum_{p=1}^{P} \mathbf{A}^v_{p,i}} \tag{8}$$

where $W^a_o, W^a_v \in \mathbb{R}^{D \times D}$ and $W^v_o, W^v_v \in \mathbb{R}^{D \times D}$ denote the learned weights of linear projections for output and value in terms of audio and visual modalities,

Method	MUSIC-Solo			VGGSound-Instruments			VGGSound-Single		
	AP(%)	IoU@0.5(%)	AUC(%)	AP(%)	IoU@0.3(%)	AUC(%)	AP(%)	IoU@0.5(%)	AUC(%)
Attention10k [39]	–	37.2	38.7	–	28.3	26.1	–	19.2	30.6
OTS [3]	69.3	26.1	35.8	47.5	25.7	24.6	29.8	32.8	35.7
DMC [21]	–	29.1	38.0	–	26.5	25.7	–	23.9	27.6
CoarsetoFine [38]	70.7	33.6	39.8	40.2	27.2	26.5	28.2	29.1	34.8
LVS [9]	70.6	41.9	40.3	42.3	32.6	28.3	29.6	34.4	38.2
EZ-VSL [31]	71.5	45.8	41.2	43.8	38.5	30.6	31.3	38.9	39.5
Mix-and-Localize [23]	68.6	30.5	40.8	44.9	49.7	32.3	32.5	36.3	38.9
DSOL [22]	–	51.4	43.7	–	50.2	32.9	–	35.7	37.2
AVGN (ours)	**77.2**	**58.1**	**48.5**	**50.5**	**55.3**	**36.7**	**35.3**	**40.8**	**42.3**

Table 1. Quantitative results of single-source localization on MUSIC-Solo, VGGSound-Instruments, and VGGSound-Single datasets.

separately. With class-aware audio-visual representations $\{\mathbf{g}_i^a\}_{i=1}^C, \{\mathbf{g}_i^v\}_{i=1}^C$ as the inputs, we use an FC layer and sigmoid operator on each modality to predict the binary probability: $p_i^a = \text{Sigmoid}(\text{FC}(\mathbf{g}_i^a)), p_i^v = \text{Sigmoid}(\text{FC}(\mathbf{g}_i^v))$ for ith class. By applying audio-visual source classes $\{y_i\}_{i=1}^C$ as the weak supervision and combining the class-constraint loss, we formulate an audio-visual grouping loss as:

$$\mathcal{L}_{\text{group}} = \sum_{i=1}^C \{\text{CE}(\mathbf{h}_i, \mathbf{e}_i) + \text{BCE}(y_i, p_i^a) + \text{BCE}(y_i, p_i^v)\}.$$
(9)

Since multiple audio sources could be in one mixture, we use binary cross-entropy loss: $\text{BCE}(\cdot)$ for each category to handle this multi-label classification problem.

With the help of the proposed class-constrained loss, we generate category-aware audio and visual representations $\{\mathbf{g}_i^a\}_{i=1}^C, \{\mathbf{g}_i^v\}_{i=1}^C$ for audio-visual alignment. Note that global audio and visual representations for N source embeddings $\{\mathbf{g}_n^a\}_{n=1}^N, \{\mathbf{g}_n^v\}_{n=1}^N$ are chosen from C categories according to the corresponding ground-truth class. Therefore, the audio-visual similarity is calculated by max-pooling audio-visual cosine similarities of class-aware audio feature $\mathbf{g}_n^a$ and the spatial-level $\{\mathbf{f}_p^v \odot \mathbf{g}_n^v\}_{p=1}^P$ across all P locations for nth source. With this category-aware similarity, we formulate the new localization loss as:

$$\mathcal{L}_{\text{loc}} = -\frac{1}{BN}\sum_{b=1}^B\sum_{n=1}^N \log \frac{\exp\left(\frac{1}{\tau}\text{sim}(\mathbf{F}_{b,n}^a, \mathbf{F}_{b,n}^v)\right)}{\sum_{m=1}^B \exp\left(\frac{1}{\tau}\text{sim}(\mathbf{F}_{b,n}^c, \mathbf{F}_{m,n}^v)\right)}$$
(10)

where $\mathbf{F}_{b,n}^a = \mathbf{g}_n^a, \mathbf{F}_{b,n}^v = \{\mathbf{f}_p^v \odot \mathbf{g}_n^v\}_{p=1}^P$ denote the class-aware audio and visual features of nth source for bth sample in the mini-batch. The overall objective of our model is simply optimized in an end-to-end manner as:

$$\mathcal{L} = \mathcal{L}_{\text{loc}} + \mathcal{L}_{\text{group}}$$
(11)

During inference, we follow the prior work [31] and use audio-visual cosine similarity map between class-aware audio-visual representations $\mathbf{g}_n^a, \{\mathbf{f}_p^v \odot \mathbf{g}_n^v\}_{p=1}^P$ to generate nth source localization map with P locations. Similarly to previous works [9,30,31], the final localization map is generated through bilinear interpolation of the similarity map.

4. Experiments
4.1. Experimental setup

Datasets. MUSIC[1] [51] contains 448 untrimmed YouTube music videos of solos and duets from 11 instrument categories. 358 solo videos are applied for training, and 90 solo videos are for evaluation. 124 duet videos are applied for training, and 17 duet videos are for evaluation. Following the prior work [23], we use MUSIC-Solo to evaluate the performance of single-source localization, and use MUSIC-Duet for evaluating multi-source localization. VGGSound-Instruments [23] consists of 32k video clips of 10s lengths from 37 musical instruments categories, which is a subset of VGG-Sound [10], and each video only has a single instrument category label. For evaluation on multi-source localization, we follow the prior work [23], and randomly concatenate two frames, resulting in one input image with a shape of 448×224, and summarize their waveforms for the mixture. Beyond those musical datasets, we filter 150k video clips of 10s lengths from the original VGG-Sound [10], which is denoted as VGGSound-Single and includes 221 categories, such as nature, animals, vehicles, people, instruments, etc. For testing, we use the full VGG-Sound Source [9] test set with 5158 videos for single-source localization. For multi-source evaluation, we follow a similar setting as VGGSound-Instruments, by randomly concatenating two frames as the input image with a size of 448×224, and summarizing their waveforms for the mixture. This results in 5158 mixed videos, which is more challenging than only 446 videos in VGGSound-Instruments. This test set is denoted as VGGSound-Duet.

Evaluation Metrics. Following the prior work [23], we use the average precision at the pixel-wise average precision (AP), Intersection over Union (IoU), and Area Under Curve (AUC) for single-source localization. When evaluating multi-source localization, we apply the class-aware average precision (CAP), permutation-invariant average precision (PIAP), Class-aware IoU (CIoU), and Area Under Curve (AUC) for a fair compar-

---
[1]Since many videos are no longer publicly available, the used dataset is smaller than the raw MUSIC dataset. For a fair comparison, we trained all models on the same training data.

Method	MUSIC-Duet				VGGSound-Instruments				VGGSound-Duet			
	CAP(%)	PIAP(%)	CIoU@0.3(%)	AUC(%)	CAP(%)	PIAP(%)	CIoU@0.1(%)	AUC(%)	CAP(%)	PIAP(%)	CIoU@0.3(%)	AUC(%)
Attention10k [39]	–	–	21.6	19.6	–	–	52.3	11.7	–	–	11.5	15.2
OTS [3]	11.6	17.7	13.3	18.5	23.3	37.8	51.2	11.2	10.5	12.7	12.2	15.8
DMC [21]	–	–	17.5	21.1	–	–	53.7	12.5	–	–	13.8	17.1
CoarsetoFine [38]	–	–	17.6	20.6	–	–	54.2	12.9	–	–	14.7	18.5
LVS [9]	–	–	22.5	20.9	–	–	57.3	13.3	–	–	17.3	19.5
EZ-VSL [31]	–	–	24.3	21.3	–	–	60.2	14.2	–	–	20.5	20.2
Mix-and-Localize [23]	47.5	54.1	26.5	21.5	21.5	37.5	73.2	15.6	16.3	22.6	21.1	20.5
DSOL [22]	–	–	30.1	22.3	–	–	74.3	15.9	–	–	22.3	21.1
AVGN (ours)	**50.6**	**57.2**	**32.5**	**24.6**	**27.3**	**42.8**	**77.5**	**18.2**	**21.9**	**28.1**	**26.2**	**23.8**

Table 2. Quantitative results of multi-source localization on MUSIC-duet, VGGSound-Instruments, and VGGSound-Duet datasets.

ison with [23]. For the threshold of IoU and CIoU, we use IoU@0.5 and CIoU@0.3 for MUSIC-Solo and MUSIC-Duet, IoU@0.3 and CIoU@0.1 for single-source and multi-source localization on VGGSound-Instruments, IoU@0.5 and CIoU@0.3 for single-source and multi-source localization on VGGSound-Single and VGGSound-Duet.

Implementation. For input images, the resolution is resized to 224×224. For input audio, we take the log spectrograms extracted from $3s$ of audio at a sample rate of 22050Hz. We follow the prior work [31] and apply STFT to generate an input tensor of size 257×300 (257 frequency bands over 300 timesteps) using 50ms windows with a hop size of 25ms. Following previous work [9, 21, 30, 31, 38], we use the lightweight ResNet18 [18] as the audio and visual encoder, and initialize the visual model using weights pre-trained on ImageNet [11]. $D = 512$, $P = 49$ for the 7×7 spatial map from the visual encoder. The depth of self-attention transformers $\phi^a(\cdot), \phi^v(\cdot)$ is 3. The model is trained for 100 epochs using the Adam optimizer [25] with a learning rate of $1e-4$ and a batch size of 128.

4.2. Comparison to prior work

In this work, we propose a novel and effective framework for sound source localization. In order to validate the effectiveness of the proposed AVGN, we comprehensively compare it to previous single-source and multi-source baselines: 1) Attention 10k [39] (2018'CVPR): the first work on single-source localization with a two-stream architecture and attention mechanism; 2) OTS [3] (2018'ECCV): a simple baseline with audio-visual correspondence as the training objective; 3) DMC [21] (2019'CVPR): a deep multimodal clustering network with audio-visual co-occurrences to learn convolutional maps for each modality in different embedding spaces; 4) CoarsetoFine [38] (2020'ECCV): a two-stage baseline with coarse-to-fine alignment for cross-modal features; 5) DSOL [22] (2020'NeurIPS): a two-stage training framework with classes as weak supervision for category-aware sound source localization; 6) LVS [9] (2021'CVPR): a contrastive network to learn audio-visual correspondence maps with hard negative mining; 7) EZ-VSL [31] (2022'ECCV): a recent strong baseline with multiple-instance contrastive learning for single-source localization; 8) Mix-and-Localize [23] (2022'CVPR): a strong multi-source baseline using a contrastive random

walk algorithm in the graph composed of images and separated sounds as nodes.

For single-source localization, we report the quantitative comparison results in Table 1. As can be seen, we achieve the best performance in terms of all metrics for three benchmarks, compared to previous self-supervised and weakly-supervised baselines. In particular, the proposed AVGN significantly outperforms DSOL [22], the current state-of-the-art weakly-supervised baseline, by 6.7 IoU@0.5 & 4.8 AUC, 5.1 IoU@0.5 & 3.8 AUC, and 5.1 IoU@0.5 & 5.1 AUC on three datasets. Moreover, we achieve superior performance gains compared to EZ-VSL [31], the current state-of-the-art self-supervised baseline, which implies the importance of extracting category-aware semantics from audio-visual inputs as the guidance for learning audio-visual alignment discriminatively. Meanwhile, our AVGN outperforms than Mix-and-Localize [23] by a large margin, where we achieve the performance gains of 8.6 AP on MUSIC-Solo, 5.6 AP on VGGSound-Instruments, and 2.7 AP on VGGSound-Single. These significant improvements demonstrate the superiority of our method in single-source localization.

In addition, significant gains in multi-source sound localization can be observed in Table 2. Compared to Mix-and-Localize [23], the current state-of-the-art multi-source localization baseline, we achieve the results gains of 5.8 CAP, 5.3 PIAP, 4.3 CIoU@0.1, and 2.6 AUC on VGGSound-Instruments. Furthermore, when evaluated on the challenging VGGSound-Duet benchmark, the proposed approach still outperforms Mix-and-Localize [23] by 5.6 CAP, 5.5 PIAP, 5.1 CIoU@0.3, and 3.3 AUC. We also achieve highly better results against DSOL [22], the weakly-supervised baseline with two training stages. These results validate the effectiveness of our approach in learning disentangled individual source semantics from mixtures and images for multi-source localization.

In order to qualitatively evaluate the localization maps, we compare the proposed AVGN with EZ-VSL [31], Mix-and-Localize [23], and DSOL [22] on both single-source and multi-source localization in Figure 3. From comparisons, three main observations can be derived: 1) Without explicit separation objectives, EZ-VSL [31], the strong single-source baseline, performs worse on the multi-source
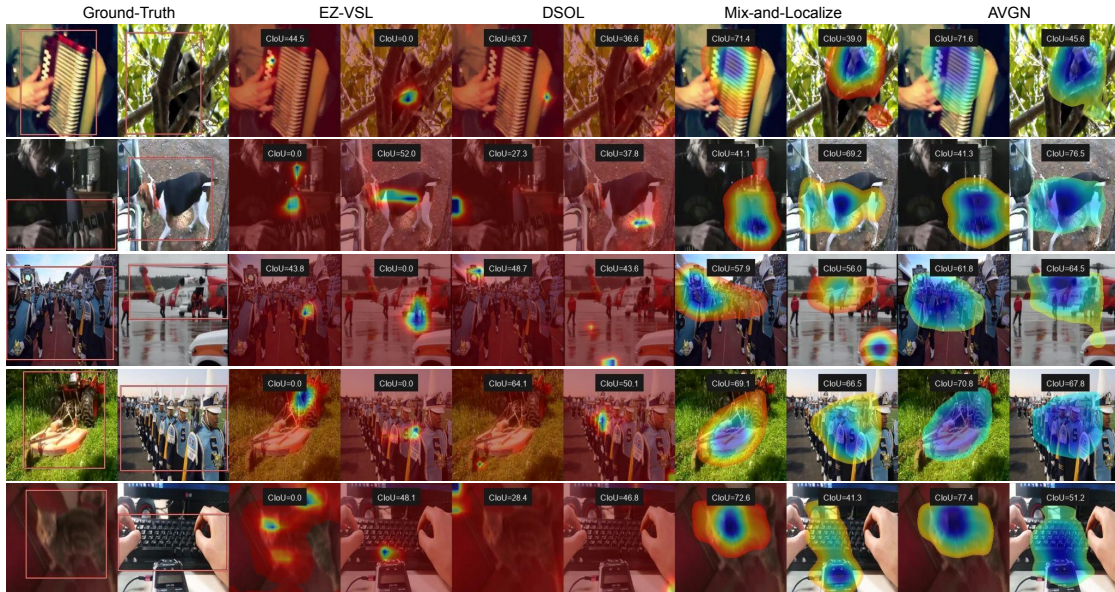
Figure 3. Qualitative comparisons with single-source and multi-source baselines on multi-source localization. Note that blue color refers to high attention values and red for low attention values. The proposed AVGN produces much more accurate and high-quality localization maps for each source.

AVCT	AVG	MUSIC-Solo			MUSIC-Duet			
		AP(%)	IoU@0.5(%)	AUC(%)	CAP(%)	PIAP(%)	CIoU@0.3(%)	AUC(%)
✗	✗	71.5	45.8	41.2	39.7	43.1	24.3	21.3
✓	✗	75.2	52.3	45.1	46.9	51.8	27.6	22.5
✗	✓	73.6	48.2	43.5	42.8	49.5	25.3	21.8
✓	✓	**77.2**	**58.1**	**48.5**	**50.6**	**57.2**	**32.5**	**24.6**

Table 3. Ablation studies on Audio-Visual Class Tokens (AVCT) and Audio-Visual Grouping (AVG).

localization. 2) the quality of localization maps generated by our method is much better than the self-supervised multi-source baseline, Mix-and-Localize [23]. 3) the proposed AVGN achieves competitive even better results on predicted maps against the weakly-supervised multi-source baseline [22] by using category labels during training. These visualizations further showcase the superiority of our simple AVGN in learning category-aware audio-visual representations to guide localization for each source.

4.3. Experimental analysis

In this section, we performed ablation studies to demonstrate the benefit of introducing the Audio-Visual Class Tokens and Audio-Visual Grouping module. We also conducted extensive experiments to explore a flexible number of sound source localization, and learned disentangled category-aware audio-visual representations.

Audio-Visual Class Tokens & Audio-Visual Grouping. In order to validate the effectiveness of the introduced audio-visual class tokens (AVCT) and audio-visual grouping (AVG), we ablate the necessity of each module and report the quantitative results in Table 3. We can observe that adding bearable AVCT to the vanilla baseline highly increases the results of single-source localization (by 3.7 AP,

6.5 IoU@0.5, and 3.9 AUC) and multi-source localization (by 7.2 CAP, 8.7 PIAP, and 3.3 CIoU@0.3), which demonstrates the benefit of category tokens in extracting disentangled high-level semantics for source localization. Meanwhile, introducing only AVG in the baseline also increases the source localization performance in terms of all metrics. More importantly, incorporating AVCT and AVG together into the baseline significantly raises the performance by 5.7 AP, 12.3 IoU@0.5 and 7.3 AUC on single-source, and by 10.9 CAP, 14.1 PIAP, 8.2 CIoU@0.3 and 3.3 AUC on multi-source localization. These improving results validate the importance of audio-visual class tokens and audio-visual grouping in extracting category-aware semantics from the mixture and image for sound localization.

Generalizing to Flexible Number of Sources. In order to show the generalizability of the proposed AVGN to a flexible number of sources, we directly transfer the model without additional training to test a mixture of 3 sources. We still achieve competitive results of 18.5 CAP, 23.7 PIAP, 22.7 CIoU@0.3, and 21.8 AUC on the challenging VGGSound-Duet dataset. These results indicate that our AVGN can support localizing a flexible number of sources from the mixture, which is different from Mix-and-Localize [23] with a
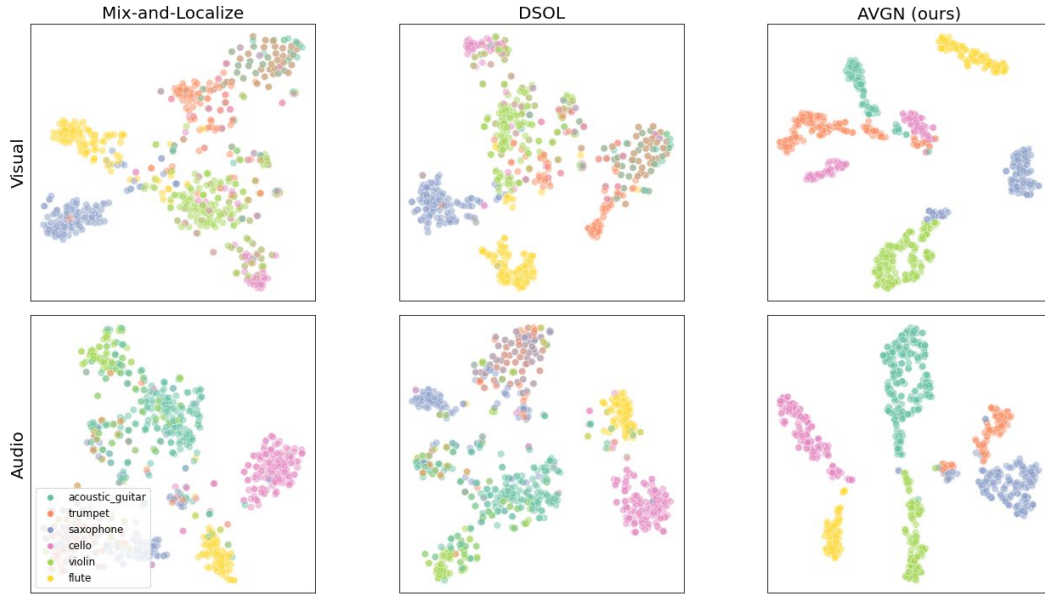
Figure 4. Qualitative comparisons of representations learned by Mix-and-Localize, DSOL, and the proposed AVGN. Note that each spot denotes the feature of one source sound, and each color refers to one source category, such as "trumpet" in orange and "cello" in pink.

fixed number of sources as the number of nodes defined in the trained contrastive random walker.

Learned Category-aware Audio-Visual Representations. Learning disentangled audio-visual representations with category-aware semantics is critical for us to localize the sound source from a mixture. To better evaluate the quality of learned category-aware features, we visualize the learned visual and audio representations of 6 categories in MUSIC-Duet by t-SNE [45], as shown in Figure 4. Note that each color refers to one category of the source sound, such as "trumpet" in orange and "cello" in pink. As can be observed in the last column, audio-visual representations extracted by the proposed AVGN are both intra-category compact and inter-category separable. In contrast to our disentangled embeddings in the audio-visual semantic space, there still exists mixtures of multiple audio-visual categories among features learned by Mix-and-Localize [23]. With the benefit of the weakly-supervised classes, DSOL [22] can extract clustered audio-visual features on some classes, such as "cello" in pink. However, most categories are mixed together as they do not incorporate the explicit audio-visual grouping mechanism in our AVGN. These meaningful visualization results further showcase the success of our AVGN in extracting compact audio-visual representations with class-aware semantics for sound source localization from the mixture.

4.4. Limitation

Although the proposed AVGN achieves superior results on both single-source and multi-source localization, the performance gains of our approach on the MUSIC-Duet benchmark with a small number of categories are not significant.

One possible reason is that our model easily overfits across the training phase, and the solution is to incorporate dropout and momentum encoders together for multi-source localization. Meanwhile, we notice that if we transfer our model to open-set source localization without additional training, it would be hard to localize unseen categories as we need to pre-define a set of categories during training and do not learn unseen category tokens to guide multi-source localization. The future work could be to add enough learnable class tokens or apply continual learning to new classes.

5. Conclusion

In this work, we present AVGN, a novel audio-visual grouping network, that can directly learn category-wise semantic features for each source from audio and visual inputs for localizing multiple sources in videos. We introduce learnable audio-visual category tokens to aggregate class-aware source features. Then, we leverage the aggregated semantic features for each source to guide localizing the corresponding regions. Compared to existing multi-source methods, our new framework can handle a flexible number of sources and learns compact audio-visual semantic representations. Empirical experiments on MUSIC, VGGSound-Instruments, and VGG-Sound Sources benchmarks demonstrate the state-of-the-art performance of our AVGN on both single-source and multi-source localization.

Broader Impact. The proposed method localizes sound sources learning from user-uploaded web videos, which might cause the model to learn internal biases in the data. For example, the model could fail to localize certain rare but crucial sound sources. These issues should be carefully addressed for the deployment of real scenarios.

References

[1] Triantafyllos Afouras, Andrew Owens, Joon Son Chung, and Andrew Zisserman. Self-supervised learning of audio-visual objects from video. In *Proceedings of European Conference on Computer Vision (ECCV)*, pages 208–224, 2020. 1, 3

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

[3] Relja Arandjelovic and Andrew Zisserman. Objects that sound. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 435–451, 2018. 5, 6

[4] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2016. 2

[5] Moitreya Chatterjee, Jonathan Le Roux, Narendra Ahuja, and Anoop Cherian. Visual scene graphs for audio source separation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1204–1213, 2021. 2

[6] Changan Chen, Unnat Jain, Carl Schissler, S. V. A. Garí, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. Soundspaces: Audio-visual navigation in 3d environments. In *Proceedings of European Conference on Computer Vision (ECCV)*, pages 17–36, 2020. 2

[7] Changan Chen, Sagnik Majumder, Al-Halah Ziad, Ruohan Gao, Santhosh Kumar Ramakrishnan, and Kristen Grauman. Learning to set waypoints for audio-visual navigation. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2021. 2

[8] Changan Chen, Carl Schissler, Sanchit Garg, Philip Kobernik, Alexander Clegg, Paul Calamia, Dhruv Batra, Philip W Robinson, and Kristen Grauman. Soundspaces 2.0: A simulation platform for visual-acoustic learning. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2022. 2

[9] Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Localizing visual sounds the hard way. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16867–16876, 2021. 1, 3, 5, 6

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

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia. Li, Kai Li, and Li Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009. 6

[12] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *arXiv preprint arXiv:1804.03619*, 2018. 2

[13] John W Fisher III, Trevor Darrell, William Freeman, and Paul Viola. Learning joint statistical models for audio-visual fusion and segregation. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2000. 3

[14] Chuang Gan, Deng Huang, Hang Zhao, Joshua B. Tenenbaum, and Antonio Torralba. Music gesture for visual sound separation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10478–10487, 2020. 2

[15] Chuang Gan, Deng Huang, Hang Zhao, Joshua B. Tenenbaum, and Antonio Torralba. Music gesture for visual sound separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[16] Ruohan Gao, Rogerio Feris, and Kristen Grauman. Learning to separate object sounds by watching unlabeled video. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 35–53, 2018. 2

[17] Ruohan Gao and Kristen Grauman. 2.5d visual sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 324–333, 2019. 2

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 6

[19] John Hershey and Michael Casey. Audio-visual sound separation via hidden markov models. *Advances in Neural Information Processing Systems*, 14, 2001. 2

[20] John Hershey and Javier Movellan. Audio vision: Using audio-visual synchrony to locate sounds. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 1999. 3

[21] Di Hu, Feiping Nie, and Xuelong Li. Deep multimodal clustering for unsupervised audiovisual learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9248–9257, 2019. 1, 2, 3, 5, 6

[22] Di Hu, Rui Qian, Minyue Jiang, Xiao Tan, Shilei Wen, Errui Ding, Weiyao Lin, and Dejing Dou. Discriminative sounding objects localization via self-supervised audiovisual matching. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, pages 10077–10087, 2020. 2, 3, 5, 6, 7, 8

[23] Xixi Hu, Ziyang Chen, and Andrew Owens. Mix and localize: Localizing sound sources in mixtures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10483–10492, 2022. 1, 2, 3, 5, 6, 7, 8

[24] Einat Kidron, Yoav Y Schechner, and Michael Elad. Pixels that sound. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005. 3

[25] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[26] Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 2

[27] Yan-Bo Lin, Yu-Jhe Li, and Yu-Chiang Frank Wang. Dual-modality seq2seq network for audio-visual event localization. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2002–2006, 2019. 2

[28] Yan-Bo Lin, Hung-Yu Tseng, Hsin-Ying Lee, Yen-Yu Lin, and Ming-Hsuan Yang. Exploring cross-video and cross-modality signals for weakly-supervised audio-visual video parsing. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2

[29] Yan-Bo Lin and Yu-Chiang Frank Wang. Audiovisual transformer with instance attention for audio-visual event localization. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 2020. 2

[30] Shentong Mo and Pedro Morgado. A closer look at weakly-supervised audio-visual source localization. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 1, 3, 5, 6

[31] Shentong Mo and Pedro Morgado. Localizing visual sounds the easy way. In *Proceedings of European Conference on Computer Vision (ECCV)*, page 218–234, 2022. 1, 3, 4, 5, 6

[32] Shentong Mo and Yapeng Tian. Multi-modal grouping network for weakly-supervised audio-visual video parsing. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2

[33] Pedro Morgado, Yi Li, and Nuno Nvasconcelos. Learning representations from audio-visual spatial alignment. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, pages 4733–4744, 2020. 2

[34] Pedro Morgado, Ishan Misra, and Nuno Vasconcelos. Robust audio-visual instance discrimination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12934–12945, 2021. 2

[35] Pedro Morgado, Nuno Nvasconcelos, Timothy Langlois, and Oliver Wang. Self-supervised generation of spatial audio for 360°video. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 2

[36] Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. Audio-visual instance discrimination with cross-modal agreement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12475–12486, June 2021. 2

[37] Andrew Owens, Jiajun Wu, Josh H. McDermott, William T. Freeman, and Antonio Torralba. Ambient sound provides supervision for visual learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 801–816, 2016. 2

[38] Rui Qian, Di Hu, Heinrich Dinkel, Mengyue Wu, Ning Xu, and Weiyao Lin. Multiple sound sources localization from coarse to fine. In *Proceedings of European Conference on Computer Vision (ECCV)*, pages 292–308, 2020. 2, 3, 5, 6

[39] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4358–4366, 2018. 1, 2, 3, 5, 6

[40] Arda Senocak, Hyeonggon Ryu, Junsik Kim, and In So Kweon. Learning sound localization better from seman-

[41] Yapeng Tian, Di Hu, and Chenliang Xu. Cyclic co-learning of sounding object visual grounding and sound separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2745–2754, 2021. 2

[42] Yapeng Tian, Dingzeyu Li, and Chenliang Xu. Unified multisensory perception: Weakly-supervised audio-visual video parsing. In *Proceedings of European Conference on Computer Vision (ECCV)*, page 436–454, 2020. 2

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

[44] Efthymios Tzinis, Scott Wisdom, Aren Jansen, Shawn Hershey, Tal Remez, Daniel PW Ellis, and John R Hershey. Into the wild with audioscope: Unsupervised audio-visual separation of on-screen sounds. *arXiv preprint arXiv:2011.01143*, 2020. 2

[45] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. 8

[46] Yu Wu and Yi Yang. Exploring heterogeneous clues for weakly-supervised audio-visual video parsing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1326–1335, 2021. 2

[47] Yu Wu, Linchao Zhu, Yan Yan, and Yi Yang. Dual attention matching for audio-visual event localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 6291–6299, 2019. 2

[48] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. *arXiv preprint arXiv:2202.11094*, 2022. 4

[49] Xudong Xu, Bo Dai, and Dahua Lin. Recursive visual sound separation using minus-plus net. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 2

[50] Hang Zhao, Chuang Gan, Wei-Chiu Ma, and Antonio Torralba. The sound of motions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1735–1744, 2019. 2

[51] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 570–586, 2018. 1, 2, 5

tically similar samples. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022. 1, 3