A Unified Audio-Visual Learning Framework for Localization, Separation, and Recognition

Shentong Mo [1] **Pedro Morgado** [2]
https://github.com/stoneMo/OneAVM

Abstract

The ability to accurately recognize, localize and separate sound sources is fundamental to any audio-visual perception task. Historically, these abilities were tackled separately, with several methods developed independently for each task. However, given the interconnected nature of source localization, separation, and recognition, independent models are likely to yield suboptimal performance as they fail to capture the interdependence between these tasks. To address this problem, we propose a unified audio-visual learning framework (dubbed OneAVM) that integrates audio and visual cues for joint localization, separation, and recognition. OneAVM comprises a *shared* audio-visual encoder and task-specific decoders trained with three objectives. The first objective aligns audio and visual representations through a localized audio-visual correspondence loss. The second tackles visual source separation using a traditional mix-and-separate framework. Finally, the third objective reinforces visual feature separation and localization by mixing images in pixel space and aligning their representations with those of all corresponding sound sources. Extensive experiments on MUSIC, VGG-Instruments, VGG-Music, and VGGSound datasets demonstrate the effectiveness of OneAVM for all three tasks, audio-visual source localization, separation, and nearest neighbor recognition, and empirically demonstrate a strong positive transfer between them.

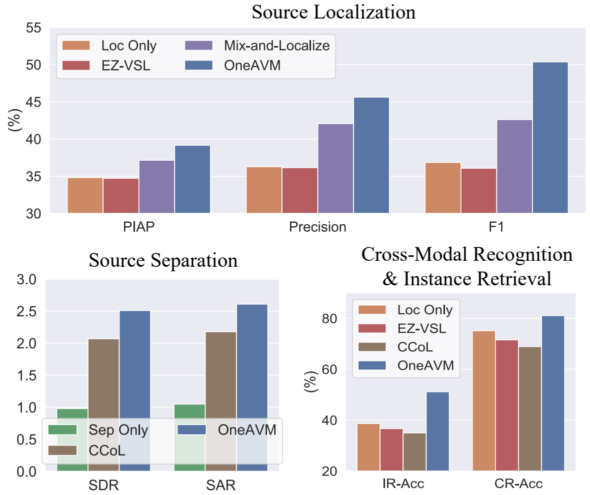

Figure 1: Sound source localization, separation, recognition (CR-ACC), and instance retrieval (IR-ACC) performance of a variety of models. The proposed framework, not only surpasses prior state-of-the-art, but it does so using a single model without requiring task-specific finetuning.

1. Introduction

Identifying, localizing, and recognizing objects or movements are critical cognitive functions, often requiring joint visual and auditory processing. In fact, a variety of neurophysiological studies have demonstrated how audio-visual interactions play a critical role in human perception. For example, humans can localize events more accurately and precisely in the presence of audio-visual sensory data compared to unisensory conditions (Odegaard et al., 2015). Cross-modal integration effects have also been identified as the cause of enhanced visual processing if reliably preceded by a sudden sound (Frassinetti et al., 2002), as well as enhanced audio-visual speech perception in noisy environments (Schwartz et al., 2004). These studies provide substantial evidence in support of the benefits of joint audio-visual processing for a wide range of tasks, from recognition to localization and source separation.

[1] Carnegie Mellon University [2] University of Wisconsin-Madison, Department of Electrical and Computer Eng. Correspondence to: Shentong Mo <shentonm@andrew.cmu.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

In the field of machine perception, researchers have also explored the potential of audio-visual interactions for each of these tasks separately. For instance, late multi-modal fusion was shown to greatly enhance object and action recognition (Kazakos et al., 2019; Xiao et al., 2020), particularly when the imbalance between dominant and non-dominant modalities is appropriately addressed. In a different line of work, two-stream neural networks, trained to match audio-visual representations, were shown capable of localizing sound sources in video data (Senocak et al., 2018; Hu et al., 2022; Mo & Morgado, 2022a;b), in a weakly supervised manner. Similarly, various multi-modal architectures have also been proposed to tackle the problem of sound source separation (Hershey & Casey, 2001; Zhao et al., 2018; Ephrat et al., 2018; Gao et al., 2018; Xu et al., 2019; Gan et al., 2020b; Tian et al., 2021).

Although promising, the aforementioned methods were developed independently for each audio-visual task and often rely on task-specific architectures and learning objectives. This not only limits the applicability of each model but also precludes positive transfer across tasks. In order to overcome these limitations and explore potential cross-task transfer, we propose a unified audio-visual learning framework capable of addressing several tasks simultaneously, including audio-visual source recognition, localization, and separation, without any task-specific fine-tuning. We dub our framework `OneAVM` where ”One” stands for the single unified model and AVM for audio-visual modeling. We utilize a two-stream encoder shared across tasks with task-specific decoders. The model is trained to learn matched audio-visual representations, which enables it to identify cross-modal associations necessary for localization. Additionally, the model is required to produce representations conducive to effective source separation through a mix-and-separate objective. We also introduce a novel mixed visual alignment objective that aligns multiple audio representations to a corresponding mixture of images. Through these learning objectives, our framework provides a more comprehensive approach to audio-visual learning, enabling effective cross-task transfer and improving the applicability of audio-visual models.

Through extensive experiments on MUSIC, VGG-Instruments, VGG-Music, and VGGSound datasets, we show the cross-task transfer benefits obtained through our unified audio-visual framework, `OneAVM`, as well as its state-of-the-art performance on visual sound localization, sound separation, and nearest neighbor recognition (see Figure 1). We highlight that, unlike the task-specific models of prior work, such capabilities are attained using a single model for all tasks. We further provide an extensive ablation study to validate the importance of simultaneous correspondence, localization, mixed audio separation, and mixed visual alignment in learning joint representations for the three downstream tasks.

In summary, this work provides three main contributions. (1) We present a novel unified audio-visual framework (`OneAVM`) capable of performing sound source localization, separation, and recognition from a single model. (2) We propose a novel mixed visual alignment objective that associates individual sound sources with mixed images. (3) Extensive experiments comprehensively demonstrate the superiority of `OneAVM` over previous baselines on various audio-visual downstream tasks.

2. Related Work

Learning Representations from Audio-Visual Correspondences. Audio-visual representation learning aims to learn joint audio-visual models that can be tuned for various downstream tasks. One extensive line of prior work (Aytar et al., 2016; Owens et al., 2016; Arandjelovic & Zisserman, 2017; Korbar et al., 2018; Senocak et al., 2018; Zhao et al., 2018; 2019; Gan et al., 2020b; Morgado et al., 2020; 2021a;b; Hershey & Casey, 2001; Ephrat et al., 2018; Hu et al., 2019; Mo et al., 2023) have addressed audio-visual representation learning by establishing the correspondence between audio and visual modalities from videos. This cross-modal alignment has proved to be beneficial for several audio-visual tasks, including audio-visual spatialization (Morgado et al., 2018; Gao & Grauman, 2019; Chen et al., 2020a; Morgado et al., 2020), event localization (Tian et al., 2018; Lin et al., 2019; Wu et al., 2019; Lin & Wang, 2020), audio-visual navigation (Chen et al., 2020a; 2021a; 2022), and parsing (Tian et al., 2020; Wu & Yang, 2021; Lin et al., 2021; Mo & Tian, 2022). In this work, we also learn from audio-visual correspondences as one of our objectives. However, we focus on integrating multiple tasks like sound source localization, separation, and recognition in a unified framework.

Visual Sound Source Localization. Visual sound source localization is a challenging task that seeks to identify objects or regions of a video corresponding to active sound sources. Early works (Hershey & Movellan, 1999; Fisher III et al., 2000; Kidron et al., 2005) used conventional machine learning techniques, *e.g.*, statistical models (Fisher III et al., 2000) and canonical correlation analysis (Kidron et al., 2005), to learn low-level alignment between audio and visual features. With the advance of deep neural nets, recent approaches (Senocak et al., 2018; Hu et al., 2019; Afouras et al., 2020; Qian et al., 2020; Chen et al., 2021b; Senocak et al., 2022; Mo & Morgado, 2022a;b; Mo & Tian, 2023a;b) apply diverse neural-net based architectures to learn from audio-visual correspondences. For example, Attention10k (Senocak et al., 2018) predicted regions of sounding objects in the image using a two-stream architecture with an attention mechanism. To improve the localization performance, LVS (Chen et al., 2021b) introduced hard

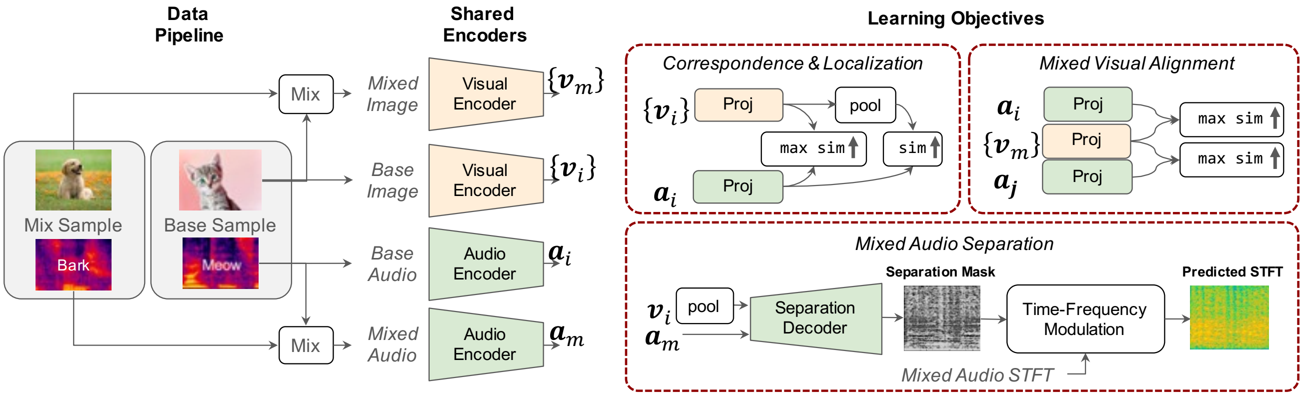

Figure 2: Illustration of the proposed audio-visual learning framework for sound source localization, separation, and recognition. First, image and audio encoders are applied to extract audio and visual features, which are trained for three separate objectives. 1) An audio-visual **correspondence and localization** objective is utilized to align corresponding audio and visual features. 2) An audio decoder is added for sound source separation using a mix-and-separate strategy (**mixed audio separation**). 3) A novel **mixed visual alignment** objective is proposed to align representations from a mixed image with the corresponding individual sound sources.

sample mining to optimize a differentiable threshold-based contrastive objective for generating discriminative audio-visual correspondence maps. Meanwhile, EZ-VSL (Mo & Morgado, 2022a) designed a multiple-instance contrastive learning loss to align regions with the most corresponding audio without negative regions involved. More recently, a contrastive random walk framework was introduced in Mix-and-Localize (Hu et al., 2022) to link each audio node with an image node using a transition probability of audio-visual similarity, which localizes individual sound sources from the mixture. However, they do not involve reconstructing separated audio and audio-visual recognition during training. Different from them, we aim to combine the sound localization objective with other audio-visual objectives, *e.g.*, separation and recognition, in a unified framework to achieve general learning for the audio-visual community.

Audio-Visual Source Separation. Audio-visual source separation aims at separating and recovering individual sound sources from an audio mixture given the image or an image region containing the sound source to be separated. In recent years, several architectures and training frameworks have been proposed to enhance audio-visual source separation. (Hershey & Casey, 2001; Zhao et al., 2018; Ephrat et al., 2018; Gao et al., 2018; Xu et al., 2019; Tian et al., 2021; Tzinis et al., 2020). For instance, Zhao *et al.* (Zhao et al., 2018) utilized the correspondence between sound sources and image regions for visually grounded separation. Other approaches, such as MP-Net (Xu et al., 2019) and CCoL (Tian et al., 2021), kept the mix-and-separate learning strategy of (Zhao et al., 2018) while improving the separation network architecture. MP-Net (Xu et al., 2019) applied a recursive MinusPlus Net to separate salient sounds from the mixture, and CCoL (Tian et al., 2021) leveraged a cyclic co-learning framework which can benefit from the visual grounding of sound sources. Similarly to CCoL, our approach also jointly optimizes for separation and localization. We show, however, superior empirical performance compared to CCoL while utilizing a much simpler architecture (without complex detection heads). Beyond visual appearance, other visual modalities have been recently explored to capture complicated visual representations, such as motion in SoM (Zhao et al., 2019), gestures through pose and keypoint detection in MG (Gan et al., 2020a), and the spatiotemporal visual scene graphs in AVSGS (Chatterjee et al., 2021). While promising, this work focuses on representations obtained from visual appearance alone (*i.e.*, RGB frames). This choice enabled us to simplify the source separation architecture and focus on developing a unified framework for audio-visual learning. A promising avenue for future research is to explore how best to incorporate such visual modalities into a unified audio-visual model.

Mixup Regularization Techniques. Mixup regularization techniques seek to create novel samples by combining existing training samples. In computer vision, popular techniques include Mix-Up (Zhang et al., 2018) and CutMix (Yun et al., 2019), which randomly combine two or more samples or image patches to create new training examples, or Cutout regularization (DeVries & Taylor, 2017) which involves randomly removing pixels or patches from the input data to force the model to learn more robust features. These regularization techniques have proven effective in improving the robustness and generalization performance of deep learning models. We found them particularly useful in our unified audio-visual learning framework. Thus, inspired by this line of work, we proposed a multi-modal variation of mix-up, where the representations of a mixed visual frame are trained to be aligned with the audio representations of all of the corresponding sound sources.

3. Method

The motivating hypothesis of this work is that audio-visual tasks like visual source localization, separation, and recognition can benefit from positive transfer across tasks. To achieve this goal, we created a unified learning framework and an audio-visual architecture to tackle these three tasks simultaneously. Our framework consists of three main components, namely correspondence & localization, as explained in section 3.2, mixed audio separation in section 3.3, and mixed visual alignment in section 3.4.

3.1. Problem Setup

Our goal is to develop a unified model for audio-visual source localization, separation, and recognition, by training on an unlabeled audio-visual dataset $\mathcal{D} = (v_i, a_i) : i = 1, \ldots, N$ leveraging the natural associations between audio and visual signals found in video data. The first task, audio-visual source localization, seeks to localize sound sources present in the audio a_i within the visual frame v_i. The second task, audio-visual source separation, is tackled through a mix-and-separate strategy (Zhao et al., 2018). The final task is recognition and instance retrieval, which aims to develop semantic representations of the data, where cross-modal association within a sample and object-level clusters across samples are readily available. To evaluate recognition and cross-modal instance retrieval, we use cosine similarity between high-level features for either retrieval or nearest neighbor classification.

Despite the disparity between the three tasks, audio-visual interactions that are useful (and easily learned) from one task might benefit the others. Therefore, to enable positive transfer across tasks, we process all audio-visual pairs through a shared encoder, regardless of the target task. Specifically, we utilize a two-stream neural network, where audio representations $\mathbf{a}_i$ are generated by an audio encoder f_a, and localized visual representations $\mathbf{v}_i^{xy}$ are computed through a visual encoder f_v. Convolutional neural networks are used for both encoders, with ResNet-18 being the chosen architecture. The audio encoder input is a log-mel spectrogram. For the visual encoder, we obtain the localized representations $\mathbf{v}_i^{xy}$ from the last feature map before global pooling. These latent representations are then fed to task-specific prediction heads, described in the following subsections.

3.2. Correspondence & Localization

Global audio-visual correspondence has been shown to lead to semantic aware representations useful for recognition tasks (Arandjelovic & Zisserman, 2017; Morgado et al., 2021b). On the other hand, localization requires an objective that promotes spatially localized audio-visual correspondence (Senocak et al., 2018; Mo & Morgado, 2022a). To accomplish both tasks, we seek both locally and globally aligned representations. Specifically, we apply two projection heads $\mathbf{a}_i^{\text{glb}} = g_a^{\text{glb}}(\mathbf{a}_i)$ and $\mathbf{a}_i^{\text{loc}} = g_a^{\text{loc}}(\mathbf{a}_i)$ to obtain two representations of audio a_i. Similarly, we project a globally pooled visual representation $\mathbf{v}_i^{\text{glb}} = g_v^{\text{glb}}(\max_{xy} \mathbf{v}_i^{xy})$, as well as a set of local visual representations $V_i^{\text{loc}} = \{g_v^{\text{loc}}(\mathbf{v}_i^{xy}) : \forall x, y\}$ of the base visual frame v_i. To align representations at both the local and global level, we define the audio-visual correspondence score between an audio a_i and video v_j as

$$s_{ij} = \mathtt{sim}(\mathbf{a}_i^{\text{glb}}, \mathbf{v}_j^{\text{glb}}) + \max_{\mathbf{v}_j^{loc} \in V_j^{loc}} \mathtt{sim}(\mathbf{a}_i^{\text{loc}}, \mathbf{v}_j^{loc}), \tag{1}$$

where $\mathtt{sim}(\cdot, \cdot)$ represents a cosine similarity between the audio and visual features. The model is then trained to optimize the average cross-modal instance discrimination loss defined as

$$\mathcal{L}_i^{\text{CL}} = -\log \frac{\exp(s_{ii}/\tau)}{\sum_{j=1}^{B} \exp(s_{ij}/\tau)} - \log \frac{\exp(s_{ii}/\tau)}{\sum_{k=1}^{B} \exp(s_{ki}/\tau)}, \tag{2}$$

where τ is a temperature hyper-parameter, and the $B - 1$ negatives are other samples in the current batch. It is worth noting that this formulation would be equivalent to the audio-visual correspondence objective of (Morgado et al., 2021a) if the similarity was computed only on global representations $s_{ij} = \mathtt{sim}(\mathbf{a}_i^{\text{glb}}, \mathbf{v}_j^{\text{glb}})$, and would be equivalent to the multiple instance contrastive learning framework of (Mo & Morgado, 2022a), if the audio-visual similarity was obtained from local representations alone, i.e. $s_{ij} = \max_{\mathbf{v}_j^{loc} \in V_j^{loc}} \mathtt{sim}(\mathbf{a}_i^{\text{loc}}, \mathbf{v}_j^{loc})$.

3.3. Mixed Audio Separation

In addition to correspondence and localization, our unified framework also tackles the cocktail party source separation problem using a mix-and-separate learning framework (Zhao et al., 2018). This involves randomly selecting two samples, (a_i, v_i) and (a_j, v_j), from the training set to create a mixed audio waveform $a_m = a_i + a_j$. An audio U-Net decoder g_{sep} is then trained to recover the waveform a_i from the audio mixture a_m, given the corresponding visual frame v_i. Specifically, the decoder receives the representations of the audio mixture $\mathbf{a}_m$ and the visual embeddings of the base sample $\mathbf{v}_i$, and applies a series of transposed convolutions and an output head to predict a time-frequency separation mask $\hat{M}_i = g_{\text{sep}}(\mathbf{a}_m, \mathbf{v}_i) \in \mathbb{R}^{T \times F}$. This separation mask is then used to modulate the input mixture STFT to separate the base audio.

$$\hat{a}_i = \textit{iSTFT}\left(\textit{STFT}(a_m) \cdot \hat{M}_i\right) \tag{3}$$

Similarly to (Zhao et al., 2018), the target masks M_i indicate the time-frequency bins in which the source is the most dominant component in the mixture, $M_i = \mathbb{1}_{|\textit{STFT}(a_i)|>|\textit{STFT}(a_m)|} \in \{0,1\}^{T\times F}$ where $\mathbb{1}$ is the indicator function applied to each *STFT* bin separately. Source separation is achieved by optimizing a binary cross-entropy loss over these binary targets M_i

$$\mathcal{L}_i^{\text{MAS}} = \frac{1}{TF}\sum_{t=1}^{T}\sum_{f=1}^{F}\text{BCE}\left(\hat{M}_i(t,f), M_i(t,f)\right). \tag{4}$$

3.4. Mixed Visual Alignment

Inspired by mixup (Zhang et al., 2018), we introduce a regularization technique called *mixed visual alignment* (MVA). Specifically, we mix visual frames from two samples, v_i and v_j, with a mix-up coefficient α to produce a mixed frame $v_m = \alpha \cdot v_i + (1-\alpha)\cdot v_j$. Then, we align the visual representation of the mixed frame v_m with the representations of both audios a_i and a_j. Let $\mathbf{a}_i^{\text{mva}} = g_a^{\text{mva}}(\mathbf{a}_i)$ and $\mathbf{a}_j^{\text{mva}} = g_a^{\text{mva}}(\mathbf{a}_j)$ be the two audio representations, and $V_m^{\text{mva}} = \{g_v^{\text{mva}}(\mathbf{v}_m^{xy}) : \forall x, y\}$ the set of local visual features for mixed frame v_m, obtained through audio and visual projection heads, $g_a^{\text{mva}}(\cdot)$ and $g_v^{\text{mva}}(\cdot)$, respectively. Mixed visual alignment is obtained by minimizing

$$\mathcal{L}_i^{\text{MVA}} = \alpha\mathcal{L}^{\text{CL}}(V_m^{\text{MVA}}, \mathbf{a}_i^{\text{MVA}}) + (1-\alpha)\mathcal{L}^{\text{CL}}(V_m^{\text{MVA}}, \mathbf{a}_j^{\text{MVA}}) \tag{5}$$

where α is the mix-up coefficient. The overall objective of our model is optimized in an end-to-end manner as:

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}(\mathcal{L}_i^{\text{CL}} + \mathcal{L}_i^{\text{MAS}} + \mathcal{L}_i^{\text{MVA}}) \tag{6}$$

We did not find it necessary to add weighing constants to the loss terms for improving downstream performance.

4. Experiments

4.1. Experimental setup

Datasets. We conducted experiments on the following audio-visual datasets. 1) *MUSIC* (Zhao et al., 2018) consists of 448 untrimmed YouTube music videos of solos and duets from 11 instrument categories. We use 358 solo videos for training and 90 solo videos for evaluation. Since some videos are no longer publicly available, the used dataset is slightly smaller than the original MUSIC dataset. For a fair comparison, we trained all models (including prior work) on the same training data. 2) *VGGSound-Instruments* (Hu et al., 2022) is a subset of VGG-Sound (Chen et al., 2020b) which includes 32k video clips of 10s lengths from 37 musical instruments categories for training and 446 videos for testing. Each video only has a single instrument category label. 3) We composed another more challenging musical subset from VGG-Sound (Chen et al., 2020b) containing 40,908 video clips from 49 music categories for training and 1201 clips for testing. We refer to this subset *VGGSound-Music*. 4) Beyond the musical datasets, we used 150k video clips from 221 categories in VGG-Sound (Chen et al., 2020b), denoted as *VGGSound-All*, where 221 classes are available in VGG-Sound Sources with source localization annotations. For testing, we used the full VGG-Sound Source (Chen et al., 2021b) test set, which contains 5158 videos with source localization annotations. 5) We also used the Kinetics-400 dataset (Carreira & Zisserman, 2017) to demonstrate the benefits of pre-training. Kinetics contains 187k video clips of human actions across 400 categories.

Evaluation Metrics. Following the prior work (Hu et al., 2022; Mo & Morgado, 2022a;b), we use the pixel-wise average precision (PIAP) from (Hu et al., 2022), as well as the Precision and F1 scores defined in (Mo & Morgado, 2022b) for visual source localization. For source separation, following (Zhao et al., 2018), we use Signal-to-Distortion Ratio (SDR) and Signal-to-Artifact Ratio (SAR). Recognition and retrieval evaluations are based on nearest-neighbors retrievals, using cosine similarity between the representations obtained from the shared encoders. We assess the accuracy of a cross-modal instance retrieval task, denoted IR-Acc, which determines how often the audio of a sample can be accurately retrieved from its visual component and vice-versa. We also assess the accuracy of both within and cross-modal nearest neighbor classifier, denoted wNN-Acc and xNN-Acc, which measures the class consistency across neighboring samples.

Implementation. The input images are resized into a 224×224 resolution. The audio is represented by log spectrograms extracted from $3s$ of audio at a sample rate of 8000Hz. We follow the prior work (Mo & Morgado, 2022a) and apply STFT to generate an input tensor of size 128×128 (128 frequency bands over 128 timesteps) using 50ms windows with a hop size of 25ms. For the audio and visual encoder, we use the ResNet18 (He et al., 2016) to extract unimodal features and initialize the visual model using weights pre-trained on ImageNet (Deng et al., 2009). Unless other specified, the decoder depth for mixed audio separation was set to 8, and the mixing coefficient for mixed visual alignment was set to $\alpha = 0.5$. For projection heads, we use one linear layer for each modality and each separate objective. The models were trained for 20 epochs using the Adam optimizer (Kingma & Ba, 2014) with a learning rate of $1e-4$ and a batch size of 128.

Table 1: **Sound source localization.** Quantitative results on VGGSound-Instruments, VGGSound-Music, VGGSound-All.

Method	VGGSound-Instruments			VGGSound-Music			VGGSound-All		
	PIAP(%)	Precision(%)	F1(%)	PIAP(%)	Precision(%)	F1(%)	PIAP(%)	Precision(%)	F1(%)
Attention10k	41.25	28.32	33.67	18.65	23.97	16.75	15.32	19.21	13.12
OTS	47.51	25.71	29.85	25.52	27.52	26.16	29.82	32.82	25.87
DMC	45.32	26.52	30.37	24.37	25.73	18.06	20.16	23.90	16.37
CoarsetoFine	40.22	27.23	32.09	26.19	28.73	27.58	28.21	29.13	21.53
DSOL	47.85	50.22	52.15	37.26	42.51	43.08	30.56	35.72	29.01
LVS	42.33	32.61	45.72	32.05	33.67	32.53	29.62	34.43	27.53
EZ-VSL	43.80	38.53	52.36	34.72	36.15	36.07	31.33	37.79	31.32
Mix-and-Localize	47.32	49.73	58.75	37.15	42.07	42.62	32.31	36.35	32.15
OneAVM (ours)	**50.67**	**55.21**	**67.26**	**39.16**	**45.63**	**50.37**	**34.52**	**39.68**	**38.75**

Table 2: **Sound source separation.** Quantitative results on MUSIC and VGGSound-Music datasets.

Method	MUSIC		VGGSound-Music	
	SDR	SAR	SDR	SAR
NMF	-0.62	2.41	-7.12	-9.01
RPCA	0.86	3.81	-5.53	-7.82
Sound-of-Pixels	4.55	10.24	0.95	1.03
MP-Net	4.82	**10.56**	1.37	1.39
CCoL	6.35	9.75	2.07	2.18
OneAVM (ours)	**7.38**	7.48	**2.51**	**2.61**

Table 3: **Nearest-neighbor recognition.** Quantitative results on VGGSound-Music dataset.

Method	IR-Acc	xNN-Acc	wNN-Acc
Sound-of-Pixels	26.92	59.09	43.64
MP-Net	30.71	64.21	49.64
CCoL	35.00	68.92	54.20
EZ-VSL	36.69	71.56	56.10
Mix-and-Localize	38.54	75.42	60.18
OneAVM (ours)	**51.17**	**81.10**	**60.91**

4.2. Comparison to prior work

We begin by comparing our unified model OneAVM to prior work on audio-visual sound source localization, separation, and recognition.

Sound source localization. To validate the effectiveness of the proposed OneAVM on sound source localization, we compare to the following prior work: 1) Attention 10k (Senocak et al., 2018) (CVPR'2018): the first baseline on sound source localization using a two-stream and attention-based neural net; 2) OTS (Arandjelovic & Zisserman, 2018) (ECCV'2018): a correspondence-based baseline for localization; 3) DMC (Hu et al., 2019) (CVPR'2019): a deep multi-modal clustering approach based on audio-visual co-occurrences; 4) CoarsetoFine (Qian et al., 2020) (ECCV'2020): a two-stage approach using coarse-to-fine embeddings alignment; 5) DSOL (Hu et al., 2020) (NeurIPS'2020): a class-based method with two-stage training; 6) LVS (Chen et al., 2021b) (CVPR'2021): a contrastive learning framework with hard negative mining to learn audio-visual correspondence maps; 7) EZ-VSL (Mo & Morgado, 2022a) (ECCV'2022): a recent weakly supervised localization framework based on multiple-instance contrastive learning; 8) Mix-and-Localize (Hu et al., 2022) (CVPR'2022): a recent method based on a contrastive random walk on a graph of images and separated sound sources.

Table 1 presents the source localization performance on VGGSound-Instruments, VGGSound-Music, and VGGSound-All datasets. The proposed OneAVM outperformed prior work on all metrics across all three datasets. We achieve significant improvements over DSOL (Hu et al., 2020), a class-supervised approach, as well as EZ-VSL (Mo & Morgado, 2022a) and Mix-and-Localize (Hu et al., 2022), two state-of-the-art weakly-supervised source localization methods. For example, on VGGSound-Instruments, we outperform the second-best method (DSOL) by 2.82 PIAP, 4.99 Precision, and 15.11 F1 score. On VGGSound-Music, the second-best method (also DSOL) was outperformed by 3.96 PIAP, 3.96 Precision, and 9.74 F1 score. Finally, the second-best method on VGGSound-All (Mix-and-Localize) was also outperformed by significant margins, 2.21 PIAP, 3.33 Precision, and 6.6 F1 score. These improvements demonstrate the effectiveness of unifying multiple tasks to learn better representations for visual sound source localization.

Sound source separation. For source separation, we compare against the following methods: 1) NMF (Virtanen, 2007): a traditional signal processing approach based on non-negative matrix factorization to generate the spectrogram of each sound source; 2) RPCA (Huang et al., 2012): a parameter-free baseline based on robust principal component analysis; 3) Sound-of-Pixels (Zhao et al., 2018): a deep learning approach that recovers separated audio conditioned on pixel-level visual features; 4) MP-Net (Xu et al., 2019): an improved audio-visual method based on recursive separation from the mixture; 5) CCoL (Tian et al., 2021):

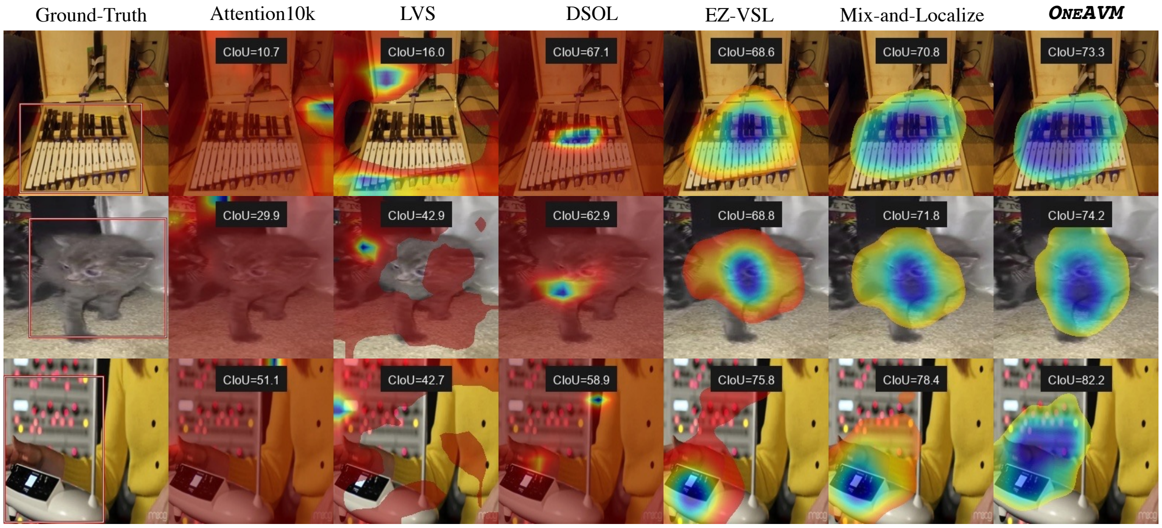

Figure 3: Example sound source localization maps. *OneAVM* produces higher-quality localization maps for each source.

Table 4: **General video pre-training on Kinetics.** Source localization, separation, and nearest neighbor recognition performance on VGGSound-Music with and without Kinetics-400 pre-training.

Train DB	**Test DB**	**Localization**			**Separation**		**Recognition**		
		PIAP	Precision	F1	SDR	SAR	IR-Acc	xNN-Acc	wNN-Acc
VGGSound-Music	VGGSound-Music	39.16	45.63	50.37	2.51	2.61	51.17	81.10	60.91
Kinetics	VGGSound-Music	22.89	26.91	38.78	0.65	0.71	8.82	9.61	22.81
Kinetics → VGGSound-Music	VGGSound-Music	**41.56**	**47.82**	**58.96**	**2.91**	**2.96**	**56.65**	**85.83**	**69.06**

a cyclic co-learning framework based on sounding object visual grounding to separate individual sound sources.

The comparison is shown in Table 2 on two datasets, MUSIC and VGGSound-Music. On the small MUSIC benchmark, we observe mixed results, with *OneAVM* outperforming all prior work by more than 1.33 SDR, while achieving a SAR score lower than other source separation methods like Sound-of-Pixels, MP-Net, and CCoL. However, on the more challenging VGGSound-Music dataset, the proposed approach outperformed all prior work both in terms of SDR and SAR. In particular, *OneAVM* outperforms methods that do not perform localization by significant margins (*e.g.* outperforming MP-Net by 1.14 SDR and 1.22 SAR) and improves over CCoL, the only other method that benefits from joint localization and source separation.

Nearest-neighbor recognition. We also compared *OneAVM* with prior work on nearest-neighbor recognition and retrieval tasks, including Sound-of-Pixels (Zhao et al., 2018), MP-Net (Xu et al., 2019), CCoL (Tian et al., 2021), EZ-VSL (Mo & Morgado, 2022a), and Mix-and-Localize (Hu et al., 2022). Table 3 shows the comparison on the VGGSound-Music dataset. The proposed approach, *OneAVM*, achieved the best performance across all metrics, outperforming the state-of-the-art localization methods like EZ-VSL and Mix-and-Localize by more than 12.6 points on cross-modal instance retrieval accuracy (IR-Acc), 5.7 points on cross-modal nearest neighbor accuracy (xNN-Acc) and 0.7 points on within-modal nearest neighbor accuracy (wNN-Acc). On the other hand, prior separation methods like CCoL, MP-Net, and Sound-of-Pixels tend to underperform in recognition tasks compared to localization methods. This outcome is not unexpected as class information is not always learnable when training exclusively for source separation. Despite its occasional usefulness, class information is not a top priority for source separation. *OneAVM* can, however, achieve superior performance on all three tasks (separation, localization, and recognition) using a single model. In fact, all results on the VGGSound-Music dataset in Tables 1, 2 and 3 were produced from the same model.

General video pre-training. All datasets considered above, either based on VGGSound or MUSIC, are composed of video samples with relatively clean audio-visual associations. However, uncurated videos tend to display weaker associations. To study the effect of uncurated videos on *OneAVM*, we utilized Kinetics-400 (which was not composed to study audio events) for pre-training and transferred the learned model to VGGSound-Music by finetuning. Source localization, separation, and recognition performance were

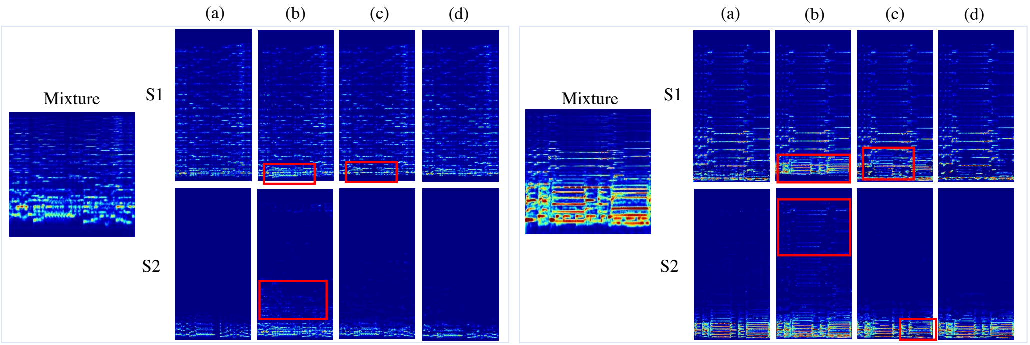

Figure 4: Qualitative visualization of sound source separation. (a) Ground-Truth; (b) Sound-of-Pixels; (c) MP-Net; (d) *OneAVM*. The proposed *OneAVM* separates each source more accurately.

Table 5: **Ablation studies.** Impact of Correspondence & Localization (CL), Mixed Audio Separation (MAS), and Mixed Visual Alignment (MVA) for visual sound localization, sound source separation, and nearest-neighbor recognition.

CL	MAS	MVA	PIAP	Precision	F1
✓			34.82	36.27	36.85
✓	✓		36.51	39.03	40.72
✓		✓	37.05	40.38	42.86
✓	✓	✓	**39.16**	**45.63**	**50.37**

(a) Visual sound localization.

CL	MAS	MVA	SDR	SAR
	✓		0.98	1.05
✓	✓		1.57	1.63
	✓	✓	1.98	2.08
✓	✓	✓	**2.51**	**2.61**

(b) Sound source separation.

CL	MAS	MVA	xIR-Acc	xNN-Acc	wNN-Acc
✓			38.68	75.14	59.49
✓	✓		38.8	76.60	60.75
✓		✓	44.47	77.43	60.54
✓	✓	✓	**51.16**	**81.10**	**60.91**

(c) Nearest-neighbor recognition.

measured on VGGSound-Music. We set the pre-training schedule to 20 epochs, and the fine-tuning schedule to 10 epochs, both with a batch size of 128.

Table 4 shows a comparison between 3 models: 1) trained on VGGSound-Music and evaluated on VGGSound-Music; 2) pre-trained on Kinetics and evaluated on VGGSound-Music; 3) pre-trained on Kinetics, finetuned on VGGSound-Music and then evaluated on VGGSound-Music. Unsurprisingly, the model trained on Kinetics alone performs worse on every task. This can be due to the domain gap between Kinetics and VGG-Sound or the weaker audio-visual associations in Kinetics videos. Nevertheless, despite the noisier samples, pre-training on Kinetics still provides a strong initialization for transfer learning. This shows that general video data, which can be more easily collected at scale, still play an important role in pre-training unified audio-visual models.

Qualitative comparisons. To further assess our unified framework, we show the localization maps and separated spectrograms generated by a single *OneAVM* model in Figures 3 and 4, respectively, and compare them with the predictions produced by specialized methods, such Attention10k (Senocak et al., 2018), LVS (Chen et al., 2021b), DSOL (Hu et al., 2020), EZ-VSL (Mo & Morgado, 2022a) for localization, and Sound-of-Pixels (Zhao et al., 2018) and MP-Net (Xu et al., 2019) for source separation. These comparisons again demonstrate the added functionality and improved performance of a unified framework like *OneAVM*.

4.3. Experimental analysis

In this section, we present the results of our ablation studies aimed at assessing the effectiveness of the various components of *OneAVM* on audio-visual source separation, localization, and recognition. Specifically, we investigate the impact of Correspondence & Localization (CL), Mixed Audio Separation (MAS), and Mixed Visual Alignment (MVA) on the performance of our approach. We also analyze the impact of the decoder depth for MAS, and the mixture ratio α in MVA, to provide insights into the optimal *OneAVM* configuration. All experiments were conducted on the VGGSound-MUSIC dataset.

CL, MAS, and MVA objectives. We assessed the effectiveness of each objective on the method's performance. Table 5 presents the model's performance on the source localization, separation, and recognition tasks. As can be seen, both source localization and recognition tasks can be significantly enhanced by adding the Mixed Audio Separation and Mixed Visual Alignment objectives. Specifically, the full *OneAVM* outperforms a model trained with Correspondence & Localization (CL) alone by 4.3 PIAP, 9.4 Precision, and 13.5 F1 score on the localization task (Table 5a), and by 12.5 IR-Acc, 6.0 xNN-Acc and 1.4 wNN-Acc on nearest neighbor recognition/retrieval (Table 5c). Sound source separation can also be enhanced significantly by adding the Correspondence & Localization, and Mixed Visual Alignment objectives, yielding a gain of 1.53 SDR and 1.6 SAR

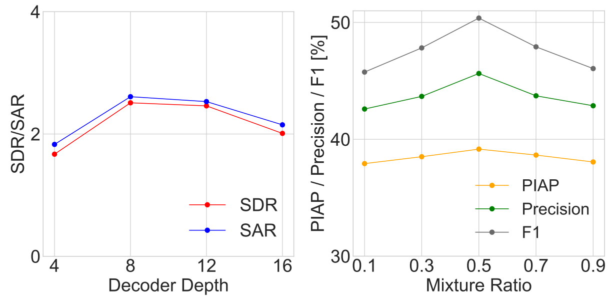

Figure 5: Impact of decoder depth in MAS on source separation and mixture ratio in MVA on source localization.

Table 6: **Localization with multiple sound sources.** Quantitative results of Precision on VGGSound-All dataset.

Method	1	2	3+
CCoL	35.31	32.25	30.51
EZ-VSL	37.79	31.67	29.32
OneAVM (ours)	**39.68**	**35.59**	**33.76**

over a method trained for separation alone (Table 5b).

These results show that localization helps separation and vice-versa, highlighting the strong interdependence of the two tasks and underscoring the importance of jointly optimizing them in our proposed audio-visual learning framework. It also shows that the proposed MVA regularization can help all tasks significantly and thus is a valuable addition to the proposed unified framework.

Decoder depth in MAS. The depth of the decoder used for source separation can affect separation performance. To assess the impact of the decoder depth, we varied it from $\{4, 8, 12, 16\}$. As shown in Figure 5, the optimal decoder depth is 8, achieving the best separation both in terms of SDR and SAR. Increasing the decoder depth beyond eight hurt performance due to overfitting.

Mixture ratio in MVA. Mixed visual alignment (MVA) can help with learning separable visual features that are aligned with multiple sound sources. The mixture ratio α is a critical hyper-parameter of MVA and can after performance significantly. To better understand the effect of the mixture ratio, we show the localization performance for varying ratios in Figure 5. OneAVM obtained optimal performance at a mixture ratio of 0.5, according to all metrics.

Localization with multiple sound sources. Lastly, we analyzed the source localization performance on samples containing multiple sound sources. Table 6 compares three methods, CCoL, EZ-VSL, and our OneAVM, for localization on the VGGSound-All dataset. EZ-VSL focuses on localization (without separation), CCoL performs joint localization and separation, and our method further trains for mixed visual alignment and global audiovisual correspondence (in addition to localization and separation). As can be seen, EZ-VSL is better than CCoL with a single object (*i.e.* when no separation is needed), but its performance drops faster as the number of objects increases. This result indicates that adding a separation objective helps localization, especially for samples with a larger number of objects. The proposed approach, OneAVM, not only outperforms EZ-VSL for videos with a single sound source, but also shows a slower (although still noticeable) decay of performance as the number of active sound sources increases.

4.4. Limitations

Although OneAVM achieves superior results on all three audio-visual downstream tasks (*i.e.* source localization, separation, and recognition), the performance gains over prior work on source separation are less consistent than those on localization. For example, our method achieves lower SAR on the MUSIC dataset than other recent methods like MP-Net. We highlight, however, both the added functionality of the model (*i.e.*, its ability to simultaneously address three tasks, as opposed to a single task) and performance improvements on other tasks when separation is used as a learning objective.

5. Conclusion

In this work, we present OneAVM, a simple yet effective approach that unifies audio-visual learning for different tasks, including localization, separation, and recognition. Specifically, we leverage correspondence and localization to align the representations of corresponding audio and video frames. We also use a mixed audio separation objective to capture discriminative audio representations from mixed audio, and introduce a mixed visual alignment objective to learn separable visual features from mixed images that can be aligned with individual sound sources. Through extensive experiments on MUSIC, VGG-Instruments, VGG-Music, and VGGSound-All datasets, we demonstrate the effectiveness of all components of our OneAVM framework and achieve favorable results in comparison to prior work on the tasks of visual sound source localization, separation, and nearest neighbor recognition.

Broader Impact. The proposed method unifies sound source localization, sound separation, and recognition 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, separate, and recognize certain rare but crucial sound sources. These issues should be carefully addressed when it comes to the deployment of real scenarios.

References

Afouras, T., Owens, A., Chung, J. S., and Zisserman, A. Self-supervised learning of audio-visual objects from video. In *Proceedings of European Conference on Computer Vision (ECCV)*, pp. 208–224, 2020. 2

Arandjelovic, R. and Zisserman, A. Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 609–617, 2017. 2, 4

Arandjelovic, R. and Zisserman, A. Objects that sound. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 435–451, 2018. 6

Aytar, Y., Vondrick, C., and Torralba, A. Soundnet: Learning sound representations from unlabeled video. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2016. 2

Carreira, J. and Zisserman, A. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6299–6308, 2017. 5

Chatterjee, M., Le Roux, J., Ahuja, N., and Cherian, A. Visual scene graphs for audio source separation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1204–1213, 2021. 3

Chen, C., Jain, U., Schissler, C., Garí, S. V. A., Al-Halah, Z., Ithapu, V. K., Robinson, P., and Grauman, K. Soundspaces: Audio-visual navigation in 3d environments. In *Proceedings of European Conference on Computer Vision (ECCV)*, pp. 17–36, 2020a. 2

Chen, C., Majumder, S., Ziad, A.-H., Gao, R., Kumar Ramakrishnan, S., and Grauman, K. Learning to set waypoints for audio-visual navigation. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2021a. 2

Chen, C., Schissler, C., Garg, S., Kobernik, P., Clegg, A., Calamia, P., Batra, D., Robinson, P. W., and Grauman, K. 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

Chen, H., Xie, W., Vedaldi, A., and Zisserman, A. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 721–725. IEEE, 2020b. 5

Chen, H., Xie, W., Afouras, T., Nagrani, A., Vedaldi, A., and Zisserman, A. Localizing visual sounds the hard way. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16867–16876, 2021b. 2, 5, 6, 8

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. ImageNet: A Large-Scale Hierarchical Image Database. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255, 2009. 5

DeVries, T. and Taylor, G. W. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 3

Ephrat, A., Mosseri, I., Lang, O., Dekel, T., Wilson, K., Hassidim, A., Freeman, W. T., and Rubinstein, M. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *arXiv preprint arXiv:1804.03619*, 2018. 2, 3

Fisher III, J. W., Darrell, T., Freeman, W., and Viola, P. Learning joint statistical models for audio-visual fusion and segregation. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2000. 2

Frassinetti, F., Bolognini, N., and Làdavas, E. Enhancement of visual perception by crossmodal visuo-auditory interaction. *Experimental brain research*, 147:332–343, 2002. 1

Gan, C., Huang, D., Zhao, H., Tenenbaum, J. B., and Torralba, A. Music gesture for visual sound separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020a. 3

Gan, C., Huang, D., Zhao, H., Tenenbaum, J. B., and Torralba, A. Music gesture for visual sound separation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10478–10487, 2020b. 2

Gao, R. and Grauman, K. 2.5d visual sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 324–333, 2019. 2

Gao, R., Feris, R., and Grauman, K. Learning to separate object sounds by watching unlabeled video. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 35–53, 2018. 2, 3

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. 5

Hershey, J. and Casey, M. Audio-visual sound separation via hidden markov models. *Advances in Neural Information Processing Systems*, 14, 2001. 2, 3

Hershey, J. and Movellan, J. Audio vision: Using audio-visual synchrony to locate sounds. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 1999. 2

Hu, D., Nie, F., and Li, X. Deep multimodal clustering for unsupervised audiovisual learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9248–9257, 2019. 2, 6

Hu, D., Qian, R., Jiang, M., Tan, X., Wen, S., Ding, E., Lin, W., and Dou, D. Discriminative sounding objects localization via self-supervised audiovisual matching. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, pp. 10077–10087, 2020. 6, 8

Hu, X., Chen, Z., and Owens, A. Mix and localize: Localizing sound sources in mixtures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10483–10492, 2022. 2, 3, 5, 6, 7

Huang, P.-S., Chen, S. D., Smaragdis, P., and Hasegawa-Johnson, M. Singing-voice separation from monaural recordings using robust principal component analysis. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 57–60, 2012. 6

Kazakos, E., Nagrani, A., Zisserman, A., and Damen, D. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5492–5501, 2019. 2

Kidron, E., Schechner, Y. Y., and Elad, M. Pixels that sound. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005. 2

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

Korbar, B., Tran, D., and Torresani, L. Cooperative learning of audio and video models from self-supervised synchronization. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 2

Lin, Y.-B. and Wang, Y.-C. F. Audiovisual transformer with instance attention for audio-visual event localization. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 2020. 2

Lin, Y.-B., Li, Y.-J., and Wang, Y.-C. F. Dual-modality seq2seq network for audio-visual event localization. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2002–2006, 2019. 2

Lin, Y.-B., Tseng, H.-Y., Lee, H.-Y., Lin, Y.-Y., and Yang, M.-H. 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

Mo, S. and Morgado, P. Localizing visual sounds the easy way. In *Proceedings of European Conference on Computer Vision (ECCV)*, pp. 218–234, 2022a. 2, 3, 4, 5, 6, 7, 8

Mo, S. and Morgado, P. A closer look at weakly-supervised audio-visual source localization. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2022b. 2, 5

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

Mo, S. and Tian, Y. Audio-visual grouping network for sound localization from mixtures. *arXiv preprint arXiv:2303.17056*, 2023a. 2

Mo, S. and Tian, Y. AV-SAM: Segment anything model meets audio-visual localization and segmentation. *arXiv preprint arXiv:2305.01836*, 2023b. 2

Mo, S., Shi, J., and Tian, Y. DiffAVA: Personalized text-to-audio generation with visual alignment. *arXiv preprint arXiv:2305.12903*, 2023. 2

Morgado, P., Nvasconcelos, N., Langlois, T., and Wang, O. Self-supervised generation of spatial audio for 360°video. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 2

Morgado, P., Li, Y., and Nvasconcelos, N. Learning representations from audio-visual spatial alignment. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, pp. 4733–4744, 2020. 2

Morgado, P., Misra, I., and Vasconcelos, N. Robust audio-visual instance discrimination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12934–12945, 2021a. 2, 4

Morgado, P., Vasconcelos, N., and Misra, I. Audio-visual instance discrimination with cross-modal agreement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12475–12486, June 2021b. 2, 4

Odegaard, B., Wozny, D. R., and Shams, L. Biases in visual, auditory, and audiovisual perception of space. *PLoS computational biology*, 11(12):e1004649, 2015. 1

Owens, A., Wu, J., McDermott, J. H., Freeman, W. T., and Torralba, A. Ambient sound provides supervision for visual learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 801–816, 2016. 2

Qian, R., Hu, D., Dinkel, H., Wu, M., Xu, N., and Lin, W. Multiple sound sources localization from coarse to fine. In *Proceedings of European Conference on Computer Vision (ECCV)*, pp. 292–308, 2020. 2, 6

Schwartz, J.-L., Berthommier, F., and Savariaux, C. Seeing to hear better: evidence for early audio-visual interactions in speech identification. *Cognition*, 93(2):B69–B78, 2004. 1

Senocak, A., Oh, T.-H., Kim, J., Yang, M.-H., and Kweon, I. S. Learning to localize sound source in visual scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4358–4366, 2018. 2, 4, 6, 8

Senocak, A., Ryu, H., Kim, J., and Kweon, I. S. Learning sound localization better from semantically similar samples. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022. 2

Tian, Y., Shi, J., Li, B., Duan, Z., and Xu, C. Audio-visual event localization in unconstrained videos. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2018. 2

Tian, Y., Li, D., and Xu, C. Unified multisensory perception: Weakly-supervised audio-visual video parsing. In *Proceedings of European Conference on Computer Vision (ECCV)*, pp. 436–454, 2020. 2

Tian, Y., Hu, D., and Xu, C. 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)*, pp. 2745–2754, 2021. 2, 3, 6, 7

Tzinis, E., Wisdom, S., Jansen, A., Hershey, S., Remez, T., Ellis, D. P., and Hershey, J. R. Into the wild with audioscope: Unsupervised audio-visual separation of on-screen sounds. *arXiv preprint arXiv:2011.01143*, 2020. 3

Virtanen, T. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(3):1066–1074, 2007. 6

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

Wu, Y., Zhu, L., Yan, Y., and Yang, Y. Dual attention matching for audio-visual event localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 6291–6299, 2019. 2

Xiao, F., Lee, Y. J., Grauman, K., Malik, J., and Feichtenhofer, C. Audiovisual slowfast networks for video recognition. *arXiv preprint arXiv:2001.08740*, 2020. 2

Xu, X., Dai, B., and Lin, D. Recursive visual sound separation using minus-plus net. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 2, 3, 6, 7, 8

Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., and Yoo, Y. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019. 3

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018. 3, 5

Zhao, H., Gan, C., Rouditchenko, A., Vondrick, C., McDermott, J., and Torralba, A. The sound of pixels. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 570–586, 2018. 2, 3, 4, 5, 6, 7, 8

Zhao, H., Gan, C., Ma, W.-C., and Torralba, A. The sound of motions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1735–1744, 2019. 2, 3