Unraveling Instance Associations: A Closer Look for Audio-Visual Segmentation

Yuanhong Chen [1] * Yuyuan Liu [1] * Hu Wang [1] Fengbei Liu [1]
Chong Wang [1] Helen Frazer[2] Gustavo Carneiro[3]
[1] Australian Institute for Machine Learning, University of Adelaide
[2] St Vincent's Hospital Melbourne
[3] Centre for Vision, Speech and Signal Processing, University of Surrey

Abstract

Audio-visual segmentation (AVS) is a challenging task that involves accurately segmenting sounding objects based on audio-visual cues. The effectiveness of audio-visual learning critically depends on achieving accurate cross-modal alignment between sound and visual objects. Successful audio-visual learning requires two essential components: 1) a challenging dataset with high-quality pixel-level multi-class annotated images associated with audio files, and 2) a model that can establish strong links between audio information and its corresponding visual object. However, these requirements are only partially addressed by current methods, with training sets containing biased audio-visual data, and models that generalise poorly beyond this biased training set. In this work, we propose a new cost-effective strategy to build challenging and relatively unbiased high-quality audio-visual segmentation benchmarks. We also propose a new informative sample mining method for audio-visual supervised contrastive learning to leverage discriminative contrastive samples to enforce cross-modal understanding. We show empirical results that demonstrate the effectiveness of our benchmark. Furthermore, experiments conducted on existing AVS datasets and on our new benchmark show that our method achieves state-of-the-art (SOTA) segmentation accuracy[1].

1. Introduction

The human nervous system exhibits multi-modal perception [44], combining input signals from different modalities to improve the detection and classification of multiple stimuli [44]. Such functionality has been emulated by recent papers [1–3, 5, 18, 32, 33] that aim to associate visual objects with their corresponding audio sequences, in a task known as audio-visual correspondence (AVC) [2, 3].

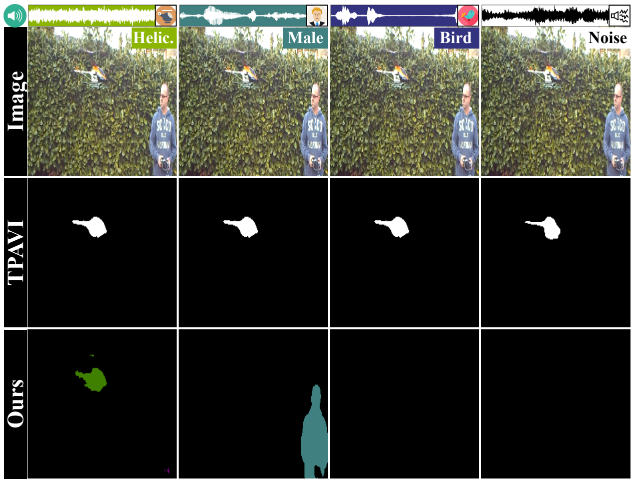

Figure 1. Current AVS datasets [54] tend to assume specific objects as consistent sound sources. Such a bias influences AVS methods, like TPAVI [54] (2nd row), to favour segmenting the presumed sound source, even when replacing the original audio with different sound types such as a person speaking (2nd column), bird chirping (3rd column), or background noise (4th row). Our paper proposes a new cost-effective strategy to build a relatively unbiased audio-visual segmentation benchmark and a supervised contrastive learning method that mines informative samples to better constrain the learning of audio-visual embeddings (last row).

A particularly interesting AVC task is the audio-visual segmentation (AVS) [54, 55] that aims to segment all pixels of the sounding visual objects using a fully supervised model. A major challenge in AVS is achieving cross-modal alignment between sound and visual objects [43]. Current datasets poorly establish and evaluate this alignment, leading to undesired system behaviour and less effective evaluation. For instance, the dataset in [54] shows a "commonsense" bias because it assumes that *certain objects are always the sound source in some scenarios*. Fig. 1 shows an example of a scene from [54] with a bias toward the segmentation of the helicopter, even though other sound sources (e.g., person's speech or bird's singing) are plau-

*First two authors contributed equally to this work.

[1] This work was supported by Australian Research Council through grant FT190100525.

sible. Such biases can reduce cross-modal alignment understanding. Another challenge in AVS datasets [54] is *the localization of the sound-producing object when multiple instances of the same object class are present*. While visual cues (e.g., motion or visual semantics) can help, some actions with silent audio can introduce false positives [34]. Spatial audio can also be helpful [40, 50], as shown in Fig. 3, where we are more likely to segment the dog on the right if the spatial audio suggests that the sounding object is located on the right-hand side of the image. Regrettably, addressing the AVS dataset challenges mentioned above by collecting new datasets can be exorbitantly expensive. Therefore, we consider alternative data collection procedures to mitigate the problems described above and to effectively enable the training and evaluation of AVS methods that can better generalise to diverse AVS conditions.

Many AVL [5, 33] and AVS [29] methods rely on audio-visual contrastive learning (AV-CL). AV-CL bears some similarities with metric learning, particularly with respect to the selection of informative samples (also called hard sampling) for improving training efficiency and model performance [38, 41, 52]. In AVC tasks, such a selection of informative training samples is not well-studied, but it is critical because the more representative visual data can overpower the weaker audio modality [39], resulting in false detections that are relatively independent of the audio-visual input, as shown in Fig. 1 for TPAVI (columns 2 to 4). Also, current AV-CL methods [29, 32] consist of unsupervised learning approaches that treat each audio-visual data pair as an independent contrastive class. However, such instance-based CL is unable to effectively mine informative samples that can mitigate the false detections mentioned above.

In this paper, we introduce a new cost-effective AVS data collection procedure for training and evaluating AVS methods that aim to mitigate the aforementioned problems of AVS datasets [54], and a new AVS method developed to address the shortcomings of AV-CL approaches. Our new AVS dataset collection and annotation, called Visual Post-production (VPO), consists of matching images from COCO [23] and audio files from VGGSound [4] based on the semantic classes of the visual objects of the images. The proposed VPO dataset has three settings: 1) the single sound-source (VPO-SS), which contains multiple visual objects, but just a single sounding object; 2) multiple sound-source (VPO-MS), which contains multiple visual objects with multiple sounding objects from different classes; and 3) multiple sound-source multi-instance (VPO-MSMI), which has multiple sets of visual objects from the same or different classes with multiple sounding objects. In these three settings, stereo sound is used to disambiguate visual objects. We also propose a new AVS method, named contrastive audio-visual pairing (CAVP), with a supervised contrastive learning approach that leverages audio-visual pairings to mine informative contrastive samples. To summarise, our main contributions are

- A new cost-effective strategy to build AVS datasets, named Visual Post-production (VPO), which aims to reduce the biases observed in current datasets [54] by pairing images [23] and audio files [4] based on the visual classes of the image objects. Three new VPO benchmarks are built using this strategy: the single sound source (VPO-SS: single sounding object per image), multiple sound sources (VPO-MS: multiple sounding objects per image from separate classes), and multiple sound sources multi-instance (VPO-MSMI: multiple sets of sounding objects from the same class).
- A new supervised audio-visual contrastive learning method that mines informative contrastive pairs from arbitrary audio-visual pairings to better constrain the learning of audio-visual embeddings.
- A thorough evaluation of SOTA AVS methods on AVSBench and VPO datasets. The methods are also assessed on AVS salient and semantically labelled objects with the resized image and traditional full-resolution setups.

We first show the effectiveness of our VPO strategy by modifying the AVSBench dataset [54] with the matching of AVSBench images with new VGGSound [4] audio files from the same classes. We then train TPAVI [54] on the original and modified AVSBench datasets and show that both datasets lead to the equivalent performance of TPAVI on the testing set of AVSBench. We also conducted experiments to test the segmentation accuracy of our proposed AVS method on AVSBench-Objects [54], AVSBench-Semantics [55], VPO-SS, VPO-MS and VPO-MSMI, and results display a consistent improvement of our method compared to the SOTA.

2. Related Works

Audio-visual Localisation (AVL) is a binary classification task for detecting sounding visual objects in videos, using image sequences and audio signals. It employs unsupervised AVL training with ImageNet pre-trained backbone models [11]. Prior research focused on creating joint audio-visual representations, with feature concatenation [2] or attention modules [5, 32, 33]. However, neglecting the contribution from audio can be a concern when audio and visual representations are not properly constrained [43]. This issue is addressed with contrastive learning [7, 8, 16], which emphasizes discriminative audio-visual feature learning for each instance [5, 18, 32, 33, 42, 46]. However, Senocak et al. [43] argue that AVL with instance discrimination may hinder genuine cross-modal semantic understanding [35]. Hence, they propose leveraging a richer positive set, obtained through strong augmentation or nearest neighbours, to facilitate cross-modal alignment. Despite these advancements, the lack of pixel-level annotation with semantic labels still hinders the accurate detection of visual objects.

Audio-visual Segmentation (AVS) addresses the limitations observed in AVL by providing pixel-level binary annotations. Zhou et al. [54] introduced the AVSBench-Object and AVSBench-Semantic benchmarks [55], which encompass single-source and multi-source AVS tasks for salient/multi-class object segmentation. The manual annotation of AVS datasets is costly and hence limits dataset diversity. Our VPO aims to mitigate this issue with off-the-shelf semantic segmentation datasets [23] combined with audio data from YouTube [4] to build datasets that facilitate model training and cross-modal alignment evaluations.

A recently published paper, developed in parallel with ours, proposes the AVS-Synthetic [27] benchmark aimed to reduce annotation costs by matching audio and visual categories to create a synthetic dataset. While our VPO and AVS-Synthetic share a similar concept of dataset creation, the data selection and partition process in AVS-Synthetic [27] has several problems, such as: 1) most annotated images consist of trivial cases containing a single sounding object, 2) it cannot assess models under different scenarios like single-source and multi-source settings, as proposed in AVSBench [54], 3) the use of binary labels limits semantic understanding, and 4) it shows ambiguous cases where multiple instances (MI) of the same class are associated with single audio. Our dataset collection addresses these four points, and in particular, it provides a cost-effective method to create spatial audio based on an object's relative position within the scene.

Like AVL, AVS methods [13, 19, 21, 24–27] use audio for segmentation queries. For example, some methods adopt MaskFormer [9] to perform image segmentation using audio queries and cross-attention layers. These methods benefit from the attention mechanism's ability to capture long-range dependencies and segment images, enhancing spatial-temporal reasoning [21] and task-related features [13, 21, 24, 25, 30, 54]. Also, certain methods [29, 30] have investigated the utilization of conditional generative models [17, 45], alongside contrastive learning, to create discriminative latent spaces. However, the reliance on binary annotations and image resizing limits their application and segmentation accuracy.

Contrastive learning has shown promise in AVL methods [5, 18, 32, 33]. These methods bring together augmented representations from the same instance as positives while separating representation pairs from different instances as negatives within a batch. The issue with current AVL contrastive learning is its reliance on self-supervision [7] to connect audio and visual representations of the same class. In our work, we propose a new supervised contrastive learning [20, 22, 48] that mines informative contrastive pairs from arbitrary audio-visual pairings to constrain the learning of audio-visual embeddings.

3. Visual Post-production (VPO) Benchmark

Our VPO benchmark includes three evaluation scenarios: 1) the single-source (VPO-SS) shown in Fig. 2 (2nd frame), 2) the multi-source (VPO-MS) displayed in Fig. 2 (3rd frame), and 3) the multi-source multi-instance (VPO-MSMI) displayed in Fig. 2 (4th frame). VPO is built by combining images and semantic segmentation masks from COCO [23] with audio files for the 21 COCO classes, including humans, animals, vehicles, sports, and electronics, sourced from VGGSound [4, 54]. The audio files were obtained from YouTube videos under the *Creative Commons* license, each trimmed to 10 seconds, as verified by [4]. We then randomly matched the COCO semantic segmentation masks with related audio files based on instance labels to create the VPO dataset. Below, we describe the three VPO settings: VPO-SS, VPO-MS, and VPO-MSMI. For dataset statistics, examples and a detailed description of the collection process please refer to *Supplementary Material*.

The **VPO-SS** comprises 12,202 samples (11,312 training and 890 testing samples), where a sample consists of an image, a pixel-level semantic segmentation mask of a single-sounding object, and an audio file of the sounding object class. During image collection, our process prioritized the inclusion of sounding visual classes from multiple categories within the same image, even when only one class matched the audio file. This strategy aims to reduce the "commonsense" bias in AVS datasets and avoid the dominance of a single visual object in the segmentation process, minimizing incidental correlations. This is done by ensuring images containing visual objects from various COCO classes are given higher priority than single-class images. By employing this collection protocol, we aim to produce a benchmark that has a diverse set of sounding visual objects because we can match images with many different types of audio files that represent visual objects, resulting in numerous combinations, as demonstrated in Fig. 2 (2nd frame).

The **VPO-MS** comprises 9,817 images, with 8,380 images for training and 1,437 for testing. Each image can include up to five sounding objects from the 21 COCO classes, where each visual object is accompanied by its pixel-level semantic segmentation mask and corresponding audio file. In total, the dataset contains 13,496 semantic segmentation masks. Following the same VPO-SS strategy, we prioritise the collection of images that have sounding objects from multiple classes but exclude images that contain multiple instances of the same class. Additionally, to prevent methods that detect all visual objects from an image from performing well in our benchmark, we randomly remove the sound of some visual objects from images containing more than two objects. We merge audio files from multiple sounding visual objects into a single file using addition operations performed on the waveform data [18, 51]. Fig. 2 (3rd frame) shows VPO-MS examples.

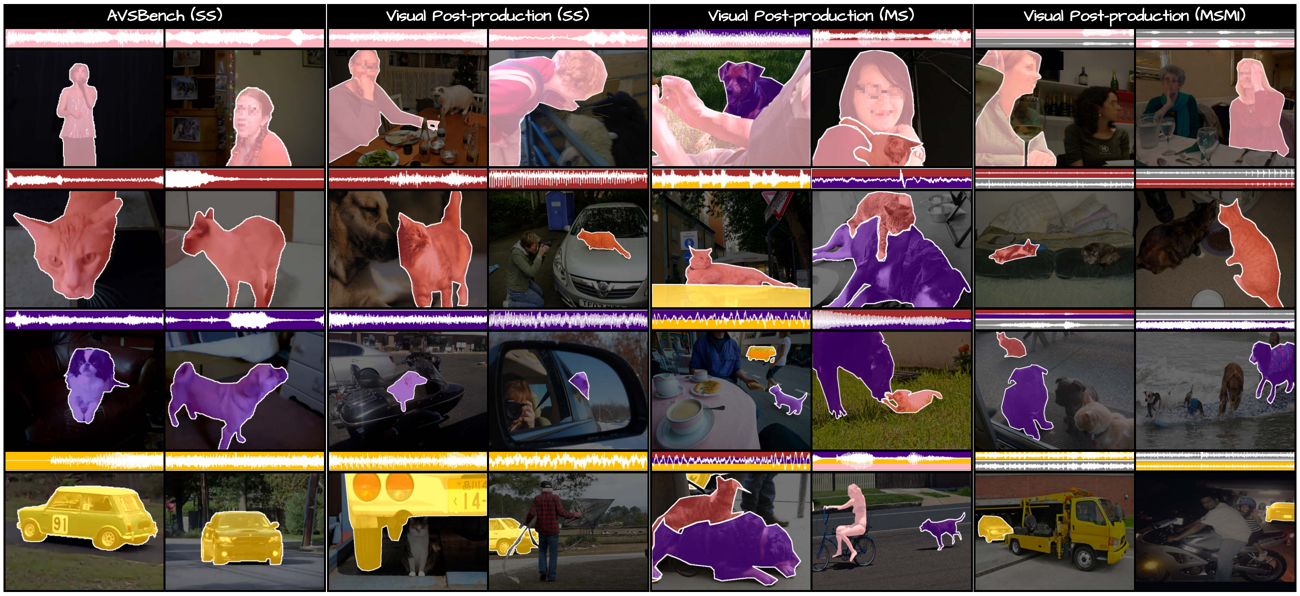
AVSBench (SS)
Visual Post-production (SS)
Visual Post-production (MS)
Visual Post-production (MSMI)

Figure 2. **VPO Benchmarks.** Using four classes, including "female", "cat", "dog", and "car", the AVSBench (SS) (1st frame) provides pixel-level multi-class annotations to the images containing a single sounding object. The proposed VPO benchmarks (2nd frame to 4th frame) pair a subset of the segmented objects in an image with relevant audio files to produce pixel-level multi-class annotations.

Based on VPO-MS, we additionally searched 3,038 images, each containing multiple instances that share a common semantic class. The resulting **VPO-MSMI** contains 12,855 images, with 11,080 images for training and 1,775 for testing. The VPO-MSMI is a challenging AVS task, represented by the accurate localisation of sound sources amid multiple sources of sound with the assistance of spatial cues. The main cue used to allow the segmentation of multiple instances in **VPO-MSMI** is the use of spatial audio to localize the sound source in an image. We simulate spatial audio with stereo sound by leveraging the object's spatial location information to modulate the volume of the left or right audio channel. Assuming we have an RGB image with resolution $H \times W$ and a particular object centred at (c_h, c_w), then the relative position of that object w.r.t the width of the image is denoted by $\alpha_i = \frac{c_w}{W}$. For instance, we show an example of processing the audio files from a human and dog in Fig. 3, utilizing position coefficients α_M, α_D for volume control and deriving c_w from ground-truth mask center of mass. Based on this modelling method, we can work with an arbitrary number of sound sources.

4. Method

The multi-class audio-visual dataset is denoted as $\mathcal{D} = \{(\mathbf{a}_i, \mathbf{x}_i, \mathbf{y}_i, \mathbf{t}_i)\}_{i=1}^{|\mathcal{D}|}$, where $\mathbf{x}_i \in \mathcal{X} \subset \mathbb{R}^{H \times W \times 3}$ is an RGB image with resolution $H \times W$, $\mathbf{a} \in \mathcal{A} \subset \mathbb{R}^{T \times F}$ denotes the Mel Spectrogram audio representation with time T and F Mel filter banks, $\mathbf{y}_i \in \mathcal{Y} \subset \{0,1\}^{H \times W \times C}$ denotes the pixel-level ground truth for the C classes (the background class is included in these C classes), and $\mathbf{t}_i \in \mathcal{Y} \subset \{0,1\}^{|C|}$ is a multi-label ground truth audio annotation.

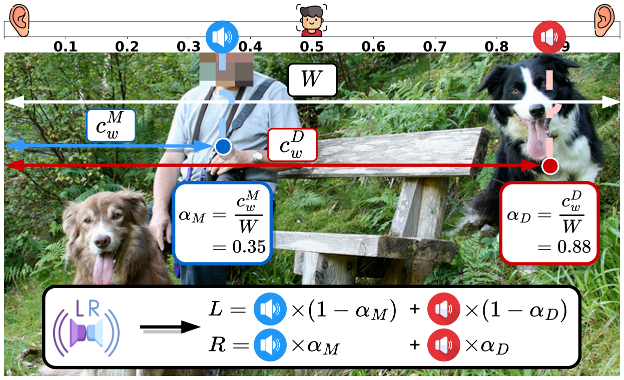
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
W
c_w^M
c_w^D
$\alpha_M = \frac{c_w^M}{W} = 0.35$
$\alpha_D = \frac{c_w^D}{W} = 0.88$
LR
$L = \times(1-\alpha_M) + \times(1-\alpha_D)$
$R = \times\alpha_M + \times\alpha_D$

Figure 3. Synthesising stereo sound for the VPO-MSMI setting.

4.1. Preliminaries about Cross-Attention

Our goal is to learn the parameters $\theta \in \Theta$ for the model $f_\theta : \mathcal{X} \times \mathcal{A} \rightarrow [0,1]^{H \times W \times C}$, which comprises the image and audio backbones that extract features with $\mathbf{u}_a = f_\gamma(\mathbf{a})$ and $\mathbf{u}_v = f_\phi(\mathbf{x})$, respectively, where $\gamma, \phi \in \theta$, and $\mathbf{u}_a, \mathbf{u}_v \in \mathcal{U}$, with $\mathcal{U}$ denoting a unified feature space. Our approach is similar to other early fusion methods [13, 36, 54] that combine audio and video features with a multi-head attention block [47] which estimates the co-occurrence of audio and visual data with $f_{MHA}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{D}}\right)\mathbf{V}$, where $\sqrt{D}$ denotes the scaling factor [47], and $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ represent the query, key and value inputs. Previous AVS methods [13, 54] usually use audio as the cross-attention (CA) query $\mathbf{Q}$ to produce $\hat{\mathbf{u}}_v = \mathbf{u}_v \oplus \mathbf{u}_v \odot f_{MHA}(\mathbf{u}_a, \mathbf{u}_v, \mathbf{u}_v)$, where $\oplus$ is the element-wise addition operator, and $\odot$ is the element-wise

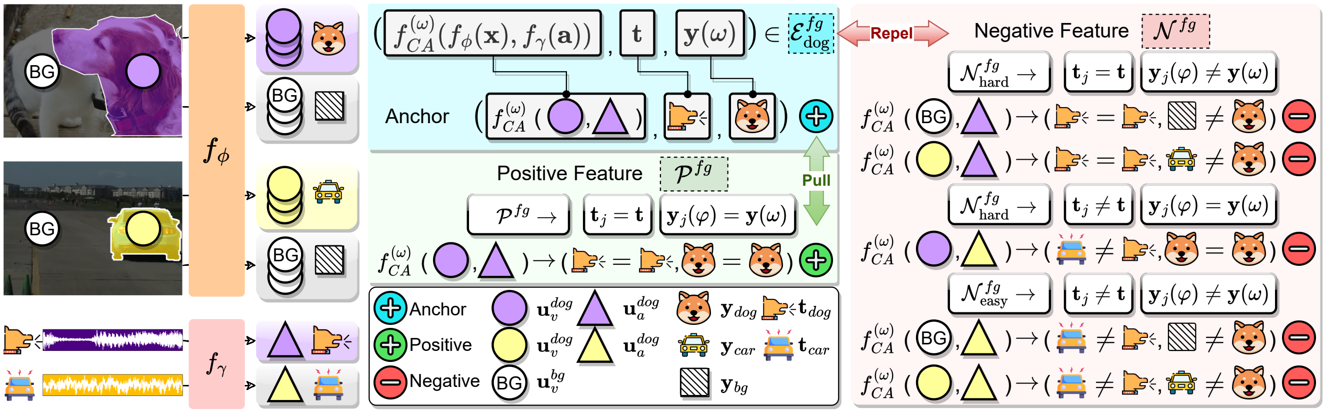
BG
BG
f_ϕ
f_γ
BG
BG
$\left(f_{CA}^{(\omega)}(f_\phi(\mathbf{x}), f_\gamma(\mathbf{a})), \mathbf{t}, \mathbf{y}(\omega) \right) \in \mathcal{E}_{dog}^{fg}$
Anchor
Repel
Negative Feature $\mathcal{N}^{fg}$
Positive Feature $\mathcal{P}^{fg}$
Pull
$\mathcal{P}^{fg} \rightarrow$ $\mathbf{t}_j = \mathbf{t}$ $\mathbf{y}_j(\varphi) = \mathbf{y}(\omega)$
$\mathcal{N}_{hard}^{fg} \rightarrow$ $\mathbf{t}_j = \mathbf{t}$ $\mathbf{y}_j(\varphi) \neq \mathbf{y}(\omega)$
$\mathcal{N}_{hard}^{fg} \rightarrow$ $\mathbf{t}_j \neq \mathbf{t}$ $\mathbf{y}_j(\varphi) = \mathbf{y}(\omega)$
$\mathcal{N}_{easy}^{fg} \rightarrow$ $\mathbf{t}_j \neq \mathbf{t}$ $\mathbf{y}_j(\varphi) \neq \mathbf{y}(\omega)$
$f_{CA}^{(\omega)}$
Anchor $\mathbf{u}_v^{dog}$ $\mathbf{u}_a^{dog}$ $\mathbf{y}_{dog}$ $\mathbf{t}_{dog}$
Positive $\mathbf{u}_v^{dog}$ $\mathbf{u}_a^{dog}$ $\mathbf{y}_{car}$ $\mathbf{t}_{car}$
Negative $\mathbf{u}_v^{bg}$ $\mathbf{y}_{bg}$

Figure 4. Illustration of our CAVP method for the "Dog" anchor. Starting with the audio-visual foreground anchor set $\mathcal{E}_{\text{dog}}^{\text{fg}}$, we create the positive and negative audio-visual features denoted by $\mathcal{P}^{\text{fg}}$ and $\mathcal{N}^{\text{fg}} = \mathcal{N}_{\text{hard}}^{\text{fg}} \cup \mathcal{N}_{\text{easy}}^{\text{fg}}$ respectively defined in Eq. (3). The CAVP loss in Eq. (4) pulls the anchor and positive audio-visual features closer while repelling the anchor and negative audio-visual features.

multiplication operator. However, we empirically observe that the use of audio as query leads to a situation where the visual feature dominates the CA process, suppressing the audio representation. We address this issue by using the visual feature as the query and removing the $\odot$ operation to produce $\hat{\mathbf{u}}_v = f_{CA}(\mathbf{u}_v, \mathbf{u}_a)$, where $f_{CA}(\mathbf{u}_v, \mathbf{u}_a) = \mathbf{u}_v \oplus f_{MHA}(\mathbf{u}_v, \mathbf{u}_a, \mathbf{u}_a)$. Another important point about $f_{MHA}(.)$ is that we replace $\text{softmax}(.)$ by $\text{sigmoid}(.)$ to enable the highlighting of multiple regions of varying sizes in the attention map that is related to the audio. Please refer to the *Supplementary Material* for the visual results.

4.2. Contrastive Audio-Visual Pairing (CAVP)

The ultimate goal of our contrastive learning is to discriminate the positive fusion of feature representations from the same semantic class and the negative fusion of feature representations from different semantic classes. Previous works [20, 28] suggest that the discrimination between samples is critical for contrastive learning, so simply drawing random negative samples from original audio-visual pairs could limit the representation learning effectiveness. To create an informative contrastive dataset, a practical way is to mine all possible combinations of audio-visual pairs in the latent space. However, it is not possible to create some of the negative pairs because we do not have access to all instance segmentation masks for each image (e.g., we cannot create $f_{CA}(\mathbf{u}_v^{cat}, \mathbf{u}_a^{dog})$ in Fig. 4 because we do not have pixels labelled as "cat"). Motivated by our VPO procedure in Sec. 3, we observe that this barrier can be overcome by randomly paring audio and visual features within the batch, as shown in Fig. 4, offering us the potential to mine most of the conceivable audio-visual combinations.

CAVP starts by dividing the original audio-visual dataset $\mathcal{D}$ into a visual dataset $\mathcal{D}^v = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{|\mathcal{D}|}$ and audio dataset $\mathcal{D}^a = \{(\mathbf{a}_k, \mathbf{t}_k)\}_{k=\mathsf{perm}(\{1,\ldots,|\mathcal{D}|\})}$, where j is a permutation of the index set of $\mathcal{D}$. We define the randomly paired audio-visual feature dataset as:

$$\mathcal{Z} = \left\{ (\mathbf{z}, \mathbf{t}, \mathbf{y}(\omega)) \,|\, \mathbf{z} = f_{CA}^{(\omega)}(f_\phi(\mathbf{x}), f_\gamma(\mathbf{a})), (\mathbf{a}, \mathbf{t}) \in \mathcal{D}^a, (\mathbf{x}, \mathbf{y}) \in \mathcal{D}^v \right\}, \tag{1}$$

where $\omega \in \Omega$ is the lattice of size $H \times W$, and $f_{CA}^{(\omega)}(.)$ is the cross-attention output at lattice position ω. To simplify the notation, we represent $\mathbf{v}(\omega) = (\mathbf{z}, \mathbf{t}, \mathbf{y}(\omega))$ and define the audio-visual anchor sets as:

$$\begin{aligned}
\mathcal{E}^{\text{fg}} &= \{\mathbf{v}(\omega) | \mathbf{v}(\omega) \in \mathcal{Z}, (\mathbf{t} = \mathbf{y}(\omega)) \wedge (\mathbf{t} \neq \text{bg}) \wedge (\mathbf{y}(\omega) \neq \text{bg})\}, \\
\mathcal{E}^{\text{unknow}} &= \{\mathbf{v}(\omega) | \mathbf{v}(\omega) \in \mathcal{Z}, (\mathbf{t}_j \neq \mathbf{y}_j(\omega)) \wedge (\mathbf{y}_j(\omega) = \text{bg})\}, \\
\mathcal{E}^{\text{bg}} &= \mathcal{Z} \setminus (\mathcal{E}^{\text{fg}} \cup \mathcal{E}^{\text{unknow}}),
\end{aligned} \tag{2}$$

where $\wedge$ and $\vee$ are respectively the "AND" and "OR" logic operators, and "bg" is the background class label. The foreground anchor set $\mathcal{E}^{\text{fg}}$ contains samples from $\mathcal{Z}$ in Eq. (1) that have the same audio and visual labels, and both are different from the background class; the $\mathcal{E}^{\text{unknow}}$ represents the anchors with uncertain semantic meaning due to the lack of instance segmentation masks for all the available targets, and the background anchor set $\mathcal{E}^{\text{bg}}$ are all the samples in $\mathcal{Z}$ that are not in $\mathcal{E}^{\text{fg}} \cup \mathcal{E}^{\text{unknow}}$.

The mining of informative contrastive samples explores all combinations of audio-visual features after cross-attention fusion to form positive sets that will enable the enhancement of the similarity between semantically related samples while diminishing the similarity of samples with different semantic concepts. For foreground anchors in $\mathcal{E}^{\text{fg}}$, its contrastive positive set $\mathcal{P}^{\text{fg}}$ and negative set $\mathcal{N}^{\text{fg}} = \mathcal{N}_{\text{hard}}^{\text{fg}} \bigcup \mathcal{N}_{\text{easy}}^{\text{fg}}$ are defined by (see Fig. 4):

$$\begin{aligned}
\mathcal{P}^{\text{fg}}(\mathbf{v}(\omega)) &= \{\mathbf{z}_j | \mathbf{v}_j(\varphi) \in \mathcal{Z}, ((\mathbf{t}_j = \mathbf{t}) \wedge (\mathbf{y}_j(\varphi) = \mathbf{y}(\omega)))\}, \\
\mathcal{N}_{\text{hard}}^{\text{fg}}(\mathbf{v}(\omega)) &= \{\mathbf{z}_j | \mathbf{v}_j(\varphi) \in \mathcal{Z}, ((\mathbf{t}_j = \mathbf{t}) \wedge (\mathbf{y}_j(\varphi) \neq \mathbf{y}(\omega))) \\
&\qquad \vee ((\mathbf{t}_j \neq \mathbf{t}) \wedge (\mathbf{y}_j(\varphi) = \mathbf{y}(\omega)))\}, \\
\mathcal{N}_{\text{easy}}^{\text{fg}}(\mathbf{v}(\omega)) &= \{\mathbf{z}_j | \mathbf{v}_j(\varphi) \in \mathcal{Z}, ((\mathbf{t}_j \neq \mathbf{t}) \wedge (\mathbf{y}_j(\varphi) \neq \mathbf{y}(\omega)))\},
\end{aligned} \tag{3}$$

Table 1. Quantitative (mIoU, F_β) audio-visual segmentation results (in %) on AVSBench dataset [54, 55] (resized to 224×224) with ResNet50 [15] backbone. Best results in **bold**, second best underlined. Improvements against the second best are in the brackets.

D-ResNet50 [15]	Method	AVSBench-Object (SS)		AVSBench-Object (MS)		AVSBench-Semantics	
		mIoU ↑	F_β ↑	mIoU ↑	F_β ↑	mIoU ↑	F_β ↑
Transformer	CATR [21]	74.80	86.60	52.80	65.30	-	-
	AuTR [26]	75.00	85.20	49.40	61.20	-	-
	AQFormer [19]	77.00	86.40	55.70	66.90	-	-
	AVSegFormer [13]	76.54	84.80	49.53	62.80	24.93	29.30
	AVSC [24]	77.02	85.24	49.58	61.51	-	-
	BAVS [25]	77.96	85.29	50.23	62.37	24.68	29.63
Per-pixel Classification	TPAVI [54]	72.79	84.80	47.88	57.80	20.18	25.20
	AVSBG [14]	74.13	85.40	44.95	56.80	-	-
	ECMVAE [30]	76.33	86.50	48.69	60.70	-	-
	DiffusionAVS [29]	75.80	86.90	49.77	62.10	-	-
	Ours	**85.77** (+7.81)	**92.86** (+5.96)	**62.39** (+6.69)	**73.62** (+6.72)	**44.70** (+19.77)	**57.76** (+28.13)

where $\varphi \in \Omega$. For background cases in $\mathcal{E}^{\text{bg}}$, the contrastive positive set is $\mathcal{P}^{\text{bg}} = \mathcal{E}^{\text{bg}}$, while the negative set is $\mathcal{N}^{\text{bg}} = \mathcal{E}^{\text{fg}}$. Let us represent the set of anchors by $\mathcal{E} = \mathcal{E}^{\text{fg}} \bigcup \mathcal{E}^{\text{bg}}$, the set of positives with $\mathcal{P}$ that is equal to $\mathcal{P}^{\text{fg}}$ if the anchor is from $\mathcal{E}^{\text{fg}}$, or equal to $\mathcal{P}^{\text{bg}}$ if the anchor is from $\mathcal{E}^{\text{bg}}$ (and similarly for $\mathcal{N}$ w.r.t. $\mathcal{N}^{\text{fg}}$ or $\mathcal{N}^{\text{bg}}$). Adopting the supervised InfoNCE [20] as the objective function to pull the anchor $\mathbf{v}(\omega) \in \mathcal{E}$ and respective positive audio-visual features closer while repelling anchors and their negative audio-visual features, we define the following loss:

$$\ell_{\text{CP}}(\mathbf{v}(\omega)) = \frac{1}{|\mathcal{P}(\mathbf{v}(\omega))|} \sum_{\mathbf{z}_p \in \mathcal{P}(\mathbf{v}(\omega))} -\log \frac{\exp\left(\mathbf{z}\cdot\mathbf{z}_p/\tau\right)}{\exp\left(\mathbf{z}\cdot\mathbf{z}_p/\tau\right) + \sum_{\mathbf{z}_n \in \mathcal{N}(\mathbf{v}(\omega))} \exp\left(\mathbf{z}\cdot\mathbf{z}_n/\tau\right)}, \tag{4}$$

where $\mathbf{z}$ is the anchor feature from $\mathbf{v}(\omega)$, and τ is the temperature hyper-parameter.

The overall training loss function is defined as:

$$\ell(\mathcal{D},\theta) = \frac{1}{|\mathcal{D}||\Omega|}\sum_{i=1}^{|\mathcal{D}|}\sum_{\varphi\in\Omega}\Big(\ell_{\text{CE}}(\mathbf{y}_i(\varphi),\hat{\mathbf{y}}_i(\varphi))\Big) + \frac{1}{|\mathcal{E}|}\sum_{\mathbf{v}(\omega)\in\mathcal{E}}\Big(\ell_{\text{CP}}(\mathbf{v}(\omega))\Big), \tag{5}$$

where $\ell_{\text{CE}}(.)$ is the cross-entropy loss, $\hat{\mathbf{y}} = f_\theta(\mathbf{x}, \mathbf{a})$ is the model (parameterised by θ) prediction, with $\hat{\mathbf{y}} \in [0,1]^{H\times W\times C}$, Ω is the image lattice, and $\ell_{\text{CP}}(.)$ is our contrastive loss, calculated based on the anchor sets $\mathcal{E}$ from Eq.(2), as specified in Eq. (4).

5. Experiments

5.1. Implementation Details

Evaluation protocols: We first adopt the widely used evaluation protocols for the AVSBench-Object [54] (including single-source (SS) and multi-source (MS) with binary labels) and AVSBench-Semantics dataset (with multi-class labels) [55] by resizing all images to 224 × 224 for fair comparison. Note that the use of binary labels and image resizing limits the application scope as well as the model performance. Hence, we follow traditional segmentation benchmarks [10, 12, 23, 37, 53] and use original image resolution for training and testing. Also following previous AVS methods [54, 55], we calculate mean intersection over union (mIoU) [12] to quantify the average segmentation quality and use F_β score with $\beta^2 = 0.3$ [31, 54] to measure precision and recall performance and false detection rate (FDR) to highlight the false positive classification in a pixel-wise manner. For training and inference details, please refer to *Supplementary Material*.

5.2. Results

Performance on resized AVSBench. We divide the performance comparison on AVSBench into the AVSBench-Objects (SS & MS) [54] and AVSBench-Semantics [55]. We compare the performance of SOTA methods with our CAVP in Tab. 1 using mIoU and F_β, which shows that our model surpasses the second-best methods w.r.t. mIoU by of 7.81% on AVSBench-Object (SS) [54], 6.69% on AVSBench-Object (MS) [54] and 19.77% on AVSBench-Semantics [55] using ResNet50 [15] backbone. Please refer to the *Supplementary Material* for results obtained with the PVT-V2-B5 [49] backbone.

Performance on original resolution AVSBench. We introduce a new benchmark based on the AVSBench-Semantics dataset [55] in Tab. 2 to address the absence of the original resolution images in previous benchmarks shown in Tab. 1. In this new benchmark, we use AVSegFormer [13][2] and TPAVI [54][3] as the baseline methods[4]. To save training and evaluation time, we train on the entire training set of AVSBench-Semantics and test the model on AVSBench-Semantics (SS) and AVSBench-Semantics (MS) subsets, as well as the entire testing set for AVSBench-Semantics to show partitioned model performance. The results on the

[2] https://github.com/vvvb-github/AVSegFormer

[3] https://github.com/OpenNLPLab/AVSBench

[4] We cannot include other methods from Tab. 1 as they were not publicly available online at the time of submission.

Table 2. Quantitative (mIoU, F_β, FDR) audio-visual segmentation results (in %) on AVSBench-Semantic dataset [55] (original resolution) with ResNet50 [15] backbone. Best results in **bold**, second best underlined. Improvements against the second best are in the last row.

D-ResNet50 [15]	Method	AVSBench-Semantics (SS)			AVSBench-Semantics (MS)			AVSBench-Semantics		
		mIoU ↑	F_β ↑	FDR ↓	mIoU ↑	F_β ↑	FDR ↓	mIoU ↑	F_β ↑	FDR ↓
Transformer	AVSegFormer [13]	68.95	82.64	15.45	33.23	45.63	43.16	41.48	56.21	38.77
Per-pixel Classification	TPAVI [54]	64.30	81.06	14.81	36.29	50.36	40.61	43.39	59.24	34.66
	Ours	**73.08**	**85.57**	**12.64**	**46.40**	**60.25**	**32.11**	**50.75**	**64.57**	**32.25**
Improvement	**Ours**	**+4.13**	**+2.93**	**-2.17**	**+10.11**	**+9.89**	**-8.50**	**+7.36**	**+5.33**	**-2.41**

Table 3. Quantitative (mIoU, F_β, FDR) audio-visual segmentation results (in %) on VPO dataset (original resolution) with ResNet50 [15] backbone. Best results in **bold**, second best underlined. Improvements against the second best are in the last row.

D-ResNet50 [15]	Method	VPO (SS)			VPO (MS)			VPO (MSMI)		
		mIoU ↑	F_β ↑	FDR ↓	mIoU ↑	F_β ↑	FDR ↓	mIoU ↑	F_β ↑	FDR ↓
Transformer	AVSegFormer [13]	57.55	73.03	19.76	58.33	74.28	22.13	54.22	70.39	25.51
Per-pixel Classification	TPAVI [54]	52.75	69.54	22.83	54.30	71.95	22.45	51.73	68.85	26.75
	Ours	**62.31**	**78.46**	**13.56**	**64.31**	**78.92**	**18.67**	**60.36**	**75.60**	**22.12**
Improvements	**Ours**	**+4.76**	**+5.43**	**-6.20**	**+5.98**	**+4.64**	**-3.46**	**+6.14**	**+5.21**	**-3.39**

Table 4. Ablation study of the CAVP components.

Method	AVSBench-Semantics		VPO (MSMI)	
	mIoU ↑	F_β ↑	mIoU ↑	F_β ↑
TPAVI [54]	42.74	58.11	52.83	70.39
$f_{CA}(\cdot)$	47.18	61.74	58.50	73.19
SupCon	48.72	61.85	59.35	74.20
CAVP	50.75	62.31	64.31	64.13

entire AVSBench-Semantics in the last columns of Tables 1 and 2 show a significant mIoU improvement of +16.55% for AVSegFromer [13], +23.21% for TPAVI, and +6.05% for our method when compared to the results obtained with low image resolution under ResNet-50 backbone [15]. These results suggest the importance of using the original resolution in the AVSBench-Semantics dataset [55]. Also in Tab. 2, results show that our method consistently outperforms the baselines by a minimum of 4.13% and 2.93% improvements in mIoU and F_β, respectively. We show a visualisation of a 6-second video clip in Fig. 5 that displays a qualitative comparison between TPAVI, AVSegFormer and our CAVP. Notice how our method approximates the ground truth segmentation of the sounding objects more consistently than the other methods. Please refer to the *Supplementary Material* for more qualitative results.

Effectiveness of Visual Post-production (VPO): To test the effectiveness of our VPO approach to build a benchmark dataset, we simulate VPO on the initial single source AVSBench-Object [54]. This involves the substitution of the original audio waveform data with samples from the same category taken from VGGSound [4] (emulating the way we build the VPO datasets). We denote this dataset as AVS-M-SS. We directly use the original TPAVI code [54] and run the model training and testing three times for both the modified and the original datasets. On AVSBench-Object (SS), TPAVI achieves mIoU=72.01 ± 0.7% and $F_\beta = 83.36 \pm 1.2\%$ (mean ± standard deviation), and on AVS-M-SS, TPAVI achieves mIoU=72.07 ± 0.7% and $F_\beta = 83.70 \pm 1.0\%$. The p-values (two-sided t-test) of mIoU and F_β are 0.66 and 0.73, respectively, which means that we fail to reject the null hypnosis that the performance on the VPO dataset is significantly different from the original one. This experiment shows the validity of building AVS datasets using pairs of audio-visual data obtained by matching images and audio based on the semantic classes of the visual objects of the images.

Table 5. Training with different proportions of positive and negative samples in (4). We use D-ResNet50 [15] and DeepLabV3+ [6] as the backbone and the segmentation architecture, respectively.

Proportion (%)		AVSBench-Semantics [55]		
Pos Sample	Neg Sample	mIoU ↑	F_β ↑	FDR ↓
10%	90%	48.88	63.03	33.13
50%	50%	50.75	64.57	32.25
90%	10%	49.19	61.82	34.50

Performance on VPO: We show model performance on our VPO benchmarks in Tab. 3 with ResNet50 [15] image backbone. Our method outperforms the baseline methods on all experimental settings by a minimum of 4.76% and 3.46% for mIoU and false detection rate, respectively.

5.3. Ablation Study

We first perform **an analysis of CAVP components** on AVSBench-Semantics [55] and VPO(MSMI) in Tab. 4. Starting from a baseline consisting of TPAVI-like cross-attention (CA) fusion layer (1st row), we replace this TPAVI CA layer by our CA layer, represented by $f_{CA}(.)$ and observe an average mIoU improvement of around +4.5% on both datasets. Subsequently, by integrating the model with the supervised contrastive learning loss defined in [20], we achieve an additional mIoU improvement of around +1%, as shown in the 3rd row. The final row displays CAVP method with all components, including the selection of informative samples, which reaches a further average mIoU improvement of +3.4%.

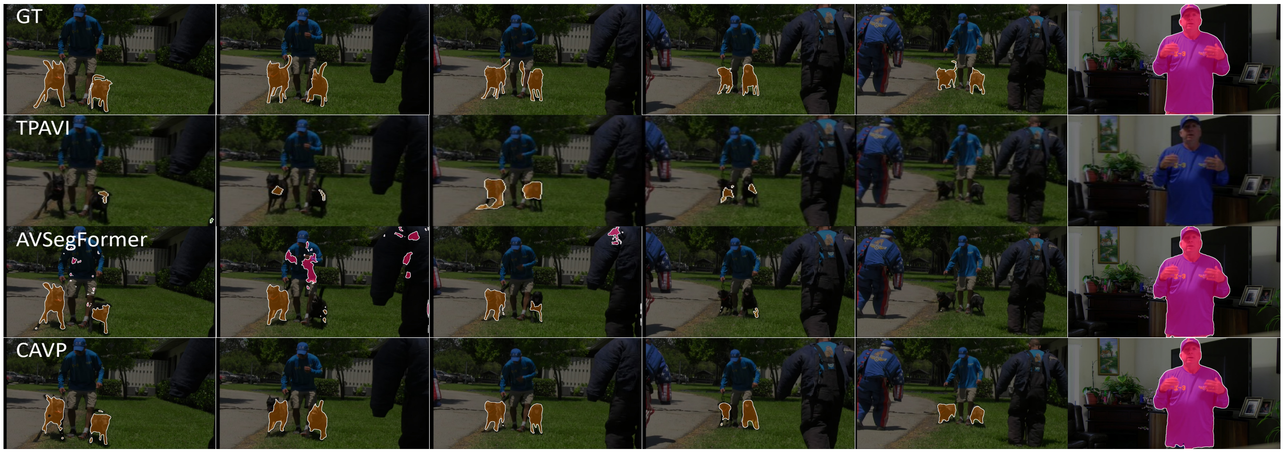

Figure 5. Qualitative audio-visual segmentation results on AVSBench-Semantics [55] by TPAVI [54], AVSegFormer [13], and our CAVP, which can be compared with the ground truth (GT) of the first row.

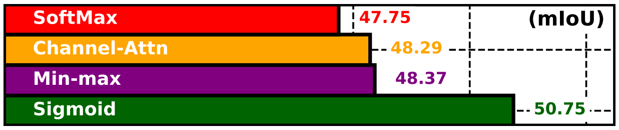

Figure 6. mIoU results for Activation functions used by $f_{CA}(.)$ on AVSBench-Semantic [55] with D-ResNet50 [15] backbone.

Cross-attention. From Sec. 4.1, recall that $f_{CA}(.)$ can have different activation functions, and we opted for the sigmoid activation to enable the highlighting of multiple regions of varying sizes in the attention map. We now study the use of different activation functions, namely: 1) softmax [47], 2) channel-wise attention adopted in AVSegFormer [13], 3) min-max normalisation to produce attention map, and the 4) sigmoid function. Results in Fig. 6 show that the setting with the "sigmoid" improves the second-best method, "min-max", by $+2.38\%$ mIoU.

Sampling Analysis. Naturally, the amount of negative samples is overwhelmingly larger than the number of positive samples in the anchor sets of Eq. (2), so we need to balance these two sets to enable more effective training. Please refer to the Supplementary Material for details on the balancing process. We conduct an ablation study of the model's performance under three different settings containing different proportions of positive and negative samples, as shown in Tab. 5. The results reveal that adopting a balanced positive and negative sampling strategy benefits the model's performance, with improvements in mIoU of 1.87% and 1.56% compared to scenarios where either 90% of the samples are negative or 90% are positive.

Mono and Stereo Sound. To emphasize the significance of stereo sound in the AVS task, we conducted an ablation study on the VPO dataset, toggling the stereo sound on and off, with results shown in Fig. 7. Notice that for all VPO subsets, the stereo sound can improve the performance on all evaluation measures[5]. This improvement is particularly noticeable in the MSMI split as it significantly enhances the model's performance across all three measures. For the numerical results, please refer to the *Supplementary Material*.

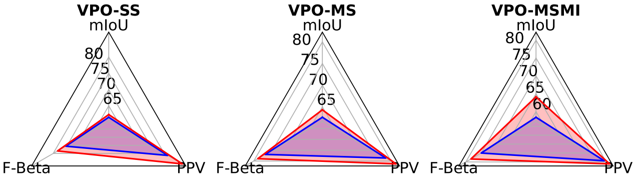

Figure 7. Mono (blue) and Stereo (red) audio VPO performance.

6. Discussion and Conclusion

In this work, we have presented new cost-effective VPO benchmarks and the innovative CAVP method for Audio-Visual Segmentation (AVS). Our proposed VPO benchmarks are both scalable, cost-effective and challenging, while our data collection and annotation protocols provide a substantial reduction of the "common-sense" bias found AVS tasks, where certain objects are always the source of sound in some scenarios. We also introduce a new supervised audio-visual contrastive learning that utilises arbitrary audio-visual pairs to mine informative contrastive pairs that can better constrain the learning of audio-visual features. Overall, our dataset and method can provide a valuable resource for future research in AVS tasks.

Limitations and future work. We recognize that our VPO dataset lacks temporal information and may exhibit a class imbalance issue similar to that observed in AVSBench-Semantics [55]. This imbalance results from our strict filtering of trivial cases with training images containing a single dominating-sounding object. Furthermore, our current approach does not support simulating spatial audio based on objects' visual depth or modelling arrival time differences in the VPO. We intend to tackle these issues in our future work to enhance AVS models' robustness and applicability.

[5] Note that PPV = 1-FDR in the graph.

References

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

[2] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE international conference on computer vision*, pages 609–617, 2017. 1, 2

[3] Relja Arandjelovic and Andrew Zisserman. Objects that sound. In *Proceedings of the European conference on computer vision (ECCV)*, pages 435–451, 2018. 1

[4] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725. IEEE, 2020. 2, 3, 7

[5] Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Localizing visual sounds the hard way. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16867–16876, 2021. 1, 2, 3

[6] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 7

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2, 3

[8] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 2

[9] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34:17864–17875, 2021. 3

[10] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 6

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2

[12] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111:98–136, 2015. 6

[13] Shengyi Gao, Zhe Chen, Guo Chen, Wenhai Wang, and Tong Lu. Avsegformer: Audio-visual segmentation with transformer. *arXiv preprint arXiv:2307.01146*, 2023. 3, 4, 6, 7, 8

[14] Dawei Hao, Yuxin Mao, Bowen He, Xiaodong Han, Yuchao Dai, and Yiran Zhong. Improving audio-visual segmentation with bidirectional generation. *arXiv preprint arXiv:2308.08288*, 2023. 6

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6, 7, 8

[16] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 2

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3

[18] 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*, pages 10483–10492, 2022. 1, 2, 3

[19] Shaofei Huang, Han Li, Yuqing Wang, Hongji Zhu, Jiao Dai, Jizhong Han, Wenge Rong, and Si Liu. Discovering sounding objects by audio queries for audio visual segmentation. *arXiv preprint arXiv:2309.09501*, 2023. 3, 6

[20] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020. 3, 5, 6, 7

[21] Kexin Li, Zongxin Yang, Lei Chen, Yi Yang, and Jun Xun. Catr: Combinatorial-dependence audio-queried transformer for audio-visual video segmentation. *arXiv preprint arXiv:2309.09709*, 2023. 3, 6

[22] Tianhong Li, Peng Cao, Yuan Yuan, Lijie Fan, Yuzhe Yang, Rogerio S Feris, Piotr Indyk, and Dina Katabi. Targeted supervised contrastive learning for long-tailed recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6918–6928, 2022. 3

[23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 2, 3, 6

[24] Chen Liu, Peike Li, Xingqun Qi, Hu Zhang, Lincheng Li, Dadong Wang, and Xin Yu. Audio-visual segmentation by exploring cross-modal mutual semantics, 2023. 3, 6

[25] Chen Liu, Peike Li, Hu Zhang, Lincheng Li, Zi Huang, Dadong Wang, and Xin Yu. Bavs: Bootstrapping audio-visual segmentation by integrating foundation knowledge. *arXiv preprint arXiv:2308.10175*, 2023. 3, 6

[26] Jinxiang Liu, Chen Ju, Chaofan Ma, Yanfeng Wang, Yu Wang, and Ya Zhang. Audio-aware query-enhanced transformer for audio-visual segmentation. *arXiv preprint arXiv:2307.13236*, 2023. 6

[27] Jinxiang Liu, Yu Wang, Chen Ju, Ya Zhang, and Weidi Xie. Annotation-free audio-visual segmentation. *arXiv preprint arXiv:2305.11019*, 2023. 3

[28] Yunze Liu, Qingnan Fan, Shanghang Zhang, Hao Dong, Thomas Funkhouser, and Li Yi. Contrastive multimodal fusion with tupleinfonce. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 754–763, 2021. 5

[29] Yuxin Mao, Jing Zhang, Mochu Xiang, Yunqiu Lv, Yiran Zhong, and Yuchao Dai. Contrastive conditional latent diffusion for audio-visual segmentation. *arXiv preprint arXiv:2307.16579*, 2023. 2, 3, 6

[30] Yuxin Mao, Jing Zhang, Mochu Xiang, Yiran Zhong, and Yuchao Dai. Multimodal variational auto-encoder based audio-visual segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 954–965, 2023. 3, 6

[31] David R Martin, Charless C Fowlkes, and Jitendra Malik. Learning to detect natural image boundaries using local brightness, color, and texture cues. *IEEE transactions on pattern analysis and machine intelligence*, 26(5):530–549, 2004. 6

[32] Shentong Mo and Pedro Morgado. A closer look at weakly-supervised audio-visual source localization. *arXiv preprint arXiv:2209.09634*, 2022. 1, 2, 3

[33] Shentong Mo and Pedro Morgado. Localizing visual sounds the easy way. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVII*, pages 218–234. Springer, 2022. 1, 2, 3

[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*, pages 12934–12945, 2021. 2

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

[36] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. *Advances in Neural Information Processing Systems*, 34:14200–14213, 2021. 4

[37] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE international conference on computer vision*, pages 4990–4999, 2017. 6

[38] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4004–4012, 2016. 2

[39] Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8238–8247, 2022. 2

[40] Kranthi Kumar Rachavarapu, Vignesh Sundaresha, AN Rajagopalan, et al. Localize to binauralize: Audio spatialization from visual sound source localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1930–1939, 2021. 2

[41] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 2

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

[43] Arda Senocak, Hyeonggon Ryu, Junsik Kim, Tae-Hyun Oh, Hanspeter Pfister, and Joon Son Chung. Sound source localization is all about cross-modal alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7777–7787, 2023. 1, 2

[44] Dana M Small and John Prescott. Odor/taste integration and the perception of flavor. *Experimental brain research*, 166: 345–357, 2005. 1

[45] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28, 2015. 3

[46] Weixuan Sun, Jiayi Zhang, Jianyuan Wang, Zheyuan Liu, Yiran Zhong, Tianpeng Feng, Yandong Guo, Yanhao Zhang, and Nick Barnes. Learning audio-visual source localization via false negative aware contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6420–6429, 2023. 2

[47] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4, 8

[48] Wenguan Wang, Tianfei Zhou, Fisher Yu, Jifeng Dai, Ender Konukoglu, and Luc Van Gool. Exploring cross-image pixel contrast for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7303–7313, 2021. 3

[49] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3):415–424, 2022. 6

[50] Xinyi Wu, Zhenyao Wu, Lili Ju, and Song Wang. Binaural audio-visual localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2961–2968, 2021. 2

[51] Yao-Yuan Yang, Moto Hira, Zhaoheng Ni, Artyom Astafurov, Caroline Chen, Christian Puhrsch, David Pollack, Dmitriy Genzel, Donny Greenberg, Edward Z Yang, et al. Torchaudio: Building blocks for audio and speech processing. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6982–6986. IEEE, 2022. 3

[52] Wenzhao Zheng, Zhaodong Chen, Jiwen Lu, and Jie Zhou. Hardness-aware deep metric learning. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 72–81, 2019. 2

[53] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127:302–321, 2019. 6

[54] Jinxing Zhou, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, and Yiran Zhong. Audio–visual segmentation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVII*, pages 386–403. Springer, 2022. 1, 2, 3, 4, 6, 7, 8

[55] Jinxing Zhou, Xuyang Shen, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, et al. Audio-visual segmentation with semantics. *arXiv preprint arXiv:2301.13190*, 2023. 1, 2, 3, 6, 7, 8