Few-Shot Incremental Learning via Foreground Aggregation and Knowledge Transfer for Audio-Visual Semantic Segmentation

Jingqiao Xiu[1], **Mengze Li**[2*], **Zongxin Yang**[3], **Wei Ji**[4*], **Yifang Yin**[5], **Roger Zimmermann**[1]

[1]National University of Singapore
[2]Hong Kong University of Science and Technology
[3]Harvard University
[4]Nanjing University
[5]Institute for Infocomm Research, A*STAR

Abstract

Audio-Visual Semantic Segmentation (AVSS) has gained significant attention in the multi-modal domain, aiming to segment video objects that produce specific sounds in the corresponding audio. Despite notable progress, existing methods still struggle to handle new classes not included in the original training set. To this end, we introduce Few-Shot Incremental Learning (FSIL) to the AVSS task, which seeks to seamlessly integrate new classes with limited incremental samples while preserving the knowledge of old classes. Two challenges arise in this new setting: (1) To reduce labeling costs, old classes within the incremental samples are treated as background, similar to silent objects. Training the model directly with background annotations may worsen the loss of distinctive knowledge about old classes, such as their outlines and sounds. (2) Most existing models adopt early cross-modal fusion with a single-tower design, incorporating more characteristics into class representations, which impedes knowledge transfer between classes based on similarity. To address these issues, we propose a **F**ew-shot **I**ncremental learning framework via class-centric foregrou**N**d aggre**G**ation and dual-tower knowl**E**dge t**R**ansfer (**FINGER**) for the AVSS task, which comprises two targeted modules: (1) The class-centric foreground aggregation gathers class-specific features for each foreground class while disregarding background features. The background class is excluded during training and inferred from the foreground predictions. (2) The dual-tower knowledge transfer postpones cross-modal fusion to separately conduct knowledge transfer for each modality. Extensive experiments validate the effectiveness of the FINGER model, significantly surpassing state-of-the-art methods.

Introduction

Audio-Visual Semantic Segmentation (AVSS) is a challenging task that has garnered significant attention in the audio-visual domain (Zhou et al. 2024). It aims to segment objects from the given video that produce specific sounds in the corresponding audio. Most existing AVSS methods (Zhou et al. 2024; Li et al. 2023a; Liu et al. 2024a; Gao et al. 2024) frame the segmentation task as pixel-level classification, where each pixel is assigned to a specific foreground class or background. These methods, trained with fixed object classes, exhibit strong performance on the predefined classes during testing. Due to the significant manpower and time costs, it is infeasible to initially label all possible sounding-body classes in the world at the pixel level for training (Cermelli et al. 2020, 2021). As user interests evolve, new classes need to be constantly incorporated into the model. For instance, as shown in Figure 1, we might begin by focusing on pop music, where instruments like the piano, cello, and guitar are prevalent. As trends shift toward classical music (with the violin as key) or rock music (centered around drums), the model must adapt to include these new classes. However, existing methods often experience a sharp performance drop when directly applied to new classes, as they lack class-

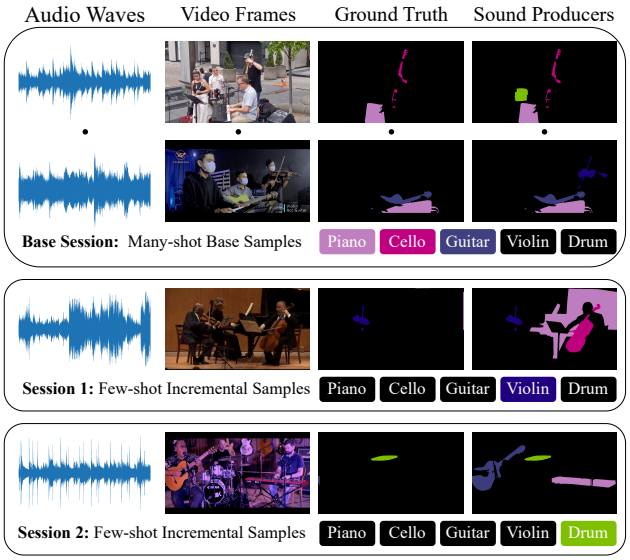


Figure 1: A toy example of training an audio-visual semantic segmentation model specialized in musical instruments. Initially, the model is trained on many-shot base samples to learn the piano, cello, and guitar classes. In subsequent sessions, the violin and drum classes are introduced to the model through few-shot incremental samples. Since multiple instruments often play together, the samples in each session typically include non-target classes, which are labeled as background in the ground truth.

*Corresponding authors.

specific knowledge. While intuitively, collecting annotated samples from these new classes could enhance the performance, the fast-paced evolution of trends makes it impractical to continually gather large-scale annotated datasets.

Imagine humans can quickly acquire new concepts with just a few examples by retaining and transferring prior knowledge. The stark contrast between human learning and neural networks motivates our exploration of few-shot incremental learning (FSIL) (Tao et al. 2020) in the AVSS task. The goal is to develop an AVSS model that can seamlessly accommodate new classes with minimal annotated samples while maintaining the prediction ability for old classes.

In pursuit of the FSIL-AVSS objective, two challenges emerge: **(1) Foreground Modeling.** The new classes introduced for incremental learning may already exist in the originally collected samples and be considered as background. This conflicts with the incremental samples and disrupts the finetuning of the AVSS model on these new classes. Moreover, to reduce labeling costs, the old classes with extensive labels in the base samples are directly treated as background in the incremental samples, without pixel-level contour annotations. If we follow the current training strategies (Zhou et al. 2024; Gao et al. 2024) and regard old classes in the incremental samples as background (similar to silent objects), the model risks forgetting important class-specific knowledge, such as sounds and contours. This seriously defeats the goal of incremental learning. **(2) Knowledge Inheritance.** The model parameters encode the knowledge learned from the base samples. Fixing these parameters hinders the integration of new knowledge from incremental samples while adjusting them threatens the retention of previously acquired information. Therefore, determining how to configure the model parameters to accommodate both old and new knowledge is a critical concern. Furthermore, with only limited samples available for new classes, the knowledge contained within these samples may be insufficient for the AVSS model to fully harness their predictive capabilities. As a result, transferring cross-modal segmentation knowledge from analogous modal information of old classes becomes essential. Existing methods typically adopt a single-tower structure (Zhou et al. 2024; Gao et al. 2024), where multi-modal information is fused early. However, this early fusion can obscure the similarities between uni-modal features across classes. For example, while dogs and cats may share morphological similarities, their vocalizations are entirely different. Prematurely merging audio and visual features before knowledge transfer may impede the mutual inspiration between classes based on feature similarities, thereby impairing the performance of incremental learning.

To tackle these issues, we introduce **FINGER**, a **F**ew-shot **I**ncremental learning framework for AVSS, featuring class-centric foregrou**N**d aggre**G**ation and dual-tower knowl**E**dge t**R**ansfer. Specifically, there are two targeted designs: **(1) Class-centric Foreground Aggregation.** During training, we introduce a proxy kernel for each foreground class to aggregate features based on their evaluated contributions to class-specific representations. This feature-level inter-class decoupling allows for exclusive training on each target class (labeled as foreground) through pixel-level binary classification, while softly masking features associated with non-target classes (labeled as background). In contrast to previous feature-centric models (Zhou et al. 2024; Gao et al. 2024) that perform multi-class classification (including background) using a single feature, our class-centric approach trains the model through binary classification on each target class feature (excluding background), which ensures that the foreground modeling remains unaffected by background annotations. During inference, pixels that do not belong to any foreground class are classified as background. **(2) Dual-tower Knowledge Transfer.** In contrast to the single-tower design of former methods (Zhou et al. 2024; Gao et al. 2024), we employ a dual-tower structure to independently achieve class-proxy aggregation and cross-class knowledge transfer for each modality, followed by a later cross-modal fusion. Additionally, directly finetuning the model on incremental samples may disrupt the knowledge memorized in the model parameters. To solve this, we freeze the parameters of the base model during incremental learning and introduce a learnable proxy calibration module, which undergoes episode training (Vinyals et al. 2016) to learn how to transfer the knowledge of old classes to new class representations.

Our contributions can be summarized as follows:

- To the best of our knowledge, we are the first to explore few-shot incremental learning in the audio-visual semantic segmentation domain, proposing a model and a benchmark to evaluate its performance.
- We introduce the proxy aggregation module and train our model without the background class to eliminate the background interference during incremental learning. In addition, we involve a dual-tower design and proxy calibration module for knowledge preservation and transfer.
- Extensive experiments demonstrate that our model consistently outperforms state-of-the-art methods in both standard and few-shot incremental benchmarks.

Related Works

Audio-Visual Semantic Segmentation (AVSS) (Zhou et al. 2024) aims to simultaneously generate pixel-level masks for sound-producing objects and predict their classes, which advances Audio-Visual Segmentation (AVS) (Mo and Tian 2023; Li et al. 2023a; Mao et al. 2023; Hao et al. 2024; Wang et al. 2024; Liu et al. 2024b; Yang et al. 2024) from binary to multi-class classification at the pixel level. Different from other visual grounding tasks (Fang et al. 2023, 2024b,a, 2025, 2024c; Ji et al. 2023c,b, 2024a, 2023a, 2024b; Xiu et al. 2024; Li et al. 2022, 2023b), it is the first pixel-level audio-visual understanding task, which is more challenging compared to instance-level audio-visual correspondence (Arandjelovic and Zisserman 2017, 2018), segment-level audio-visual event localization (Tian et al. 2018; Wu et al. 2019) and audio-visual video parsing (Tian, Li, and Xu 2020; Wu and Yang 2021; Mo and Tian 2022), and patch-level sound source localization (Chen et al. 2021; Mo and Morgado 2022). Most existing AVSS methods rely on the fusion of audio and visual features. For instance, AVSBench

(Zhou et al. 2022c) employs a temporal pixel-wise audio-visual interaction module to incorporate audio semantics as guidance for the visual segmentation process. AVSegFormer (Gao et al. 2024) introduces an audio-queried transformer decoder to focus on mixed audio-visual features of interest. Additionally, BAVS (Liu et al. 2024a) uses large foundation models to enhance audio-visual segmentation.

Few-Shot Incremental Learning (FSIL) requires the ability to quickly learn new classes with only a few samples while not forgetting previously learned classes. Feature space-based methods (Akyürek et al. 2022; Hersche et al. 2022; Yang et al. 2023; Kim et al. 2023) design the feature space to achieve more robust and efficient feature representations. In particular, some methods (Zhou et al. 2022a; Song et al. 2023) generate synthetic classes and corresponding samples to prospectively reserve embedding space for future updates. Dynamic structure-based methods (Tao et al. 2020; Zhang et al. 2021; Yang et al. 2022; Kang et al. 2023) adjust model structures or inter-class relationships as data streams continuously change, allowing for the integration of new knowledge while retaining old knowledge. Meta-learning-based methods (Zhou et al. 2022b; Chi et al. 2022) train models by simulating a series of pseudo-incremental scenarios, enabling adaptation to real incremental scenarios. Knowledge distillation-based methods (Cheraghian et al. 2021; Dong et al. 2021) optimize knowledge distillation frameworks to address class imbalance and overfitting in few-shot scenarios, facilitating knowledge transfer between sessions. However, most FSIL work focuses on simple image recognition, with only a few studies (Cermelli et al. 2021; Shi et al. 2022; Shan, Zhou, and Zhao 2023) applying it to more complex image segmentation tasks.

Methodology

Overview

Task Definition. Given an audio A and a video $V \in \mathbb{R}^{T\times 3\times H\times W}$ with T frames of resolution $H \times W$, the goal is to predict segmentation masks $M \in \mathbb{R}^{T\times H\times W}$, where each element M_{tij} corresponds to the sound-producing class of the pixel (i, j) at frame t. Pixels not associated with any sound at that moment should be classified as background.

A few-shot incremental task consists of a base session and m incremental sessions. The training and test sets are denoted by $\{D^0_{\text{train}}, D^1_{\text{train}}, \cdots, D^m_{\text{train}}\}$ and $\{D^0_{\text{test}}, D^1_{\text{test}}, \cdots, D^m_{\text{test}}\}$, respectively. Each D^s_{train} and D^s_{test} shares the same label space C^s (target classes) and includes samples containing C^s but not restricting the presence of other classes. In the training set, classes other than target classes C^s are labeled as the background, and in the test set, they are marked as the ignored class.

While the base training set contains sufficient labeled samples, the training set of each incremental session follows an $|C^s|$-way K-shot format, with $|C^s|$ classes per session and K labeled samples per class. It is worth noting that classes from different sessions do not overlap, and during session s, the model has access only to D^s_{train}. The model in session s is assessed using the test sets from the current and preceding sessions, denoted as $D^0_{\text{test}} \cup \cdots \cup D^s_{\text{test}}$.

Model Architecture. As illustrated in Figure 2, our model embraces the class-centric dual-tower architecture. The audio tower consists of an audio encoder, a proxy aggregation module, and an additional proxy calibration module for incremental sessions. The visual tower includes a visual encoder, a pixel decoder, hierarchical proxy aggregation modules, and extra calibration modules for incremental sessions. Both towers receive class proxies as part of their inputs and output to a proxy fusion module for subsequent processing.

Audio Encoder. Following previous works (Zhou et al. 2022c; Gao et al. 2024), we first resample the audio to 16 kHz mono and apply a short-time Fourier transform (STFT) to obtain the mel spectrum, which is then fed into an audio encoder to extract audio features. The audio preprocessing and encoding yield $F_a \in \mathbb{R}^{T\times D}$, where D represents the embedding dimension.

Visual Encoder. We employ a visual encoder to extract multi-scale visual features, including $1/4$, $1/8$, $1/16$, and $1/32$ of the original size.

Pixel Decoder. Features extracted by the visual encoder at scales of $1/8$, $1/16$, and $1/32$ are flattened and concatenated before being fed into the pixel decoder. The output from the pixel decoder is then split and reshaped back to the original dimensions, yielding multi-scale output features, which can be expressed as follows:

$$F_v = \{F^1_v, F^2_v, F^3_v\}, \tag{1}$$

where $F^i_v \in \mathbb{R}^{T\times D\times \frac{H}{2^{6-i}}\times \frac{W}{2^{6-i}}}$ and $i \in [1, 2, 3]$. Thereafter, the features are processed through a Feature Pyramid Network (FPN) (Lin et al. 2017). Specifically, the $1/8$-scale features are upsampled by a factor of 2 and then added to the $1/4$-scale features from the visual encoder, resulting in the mask features $F_m \in \mathbb{R}^{T\times D\times \frac{H}{4}\times \frac{W}{4}}$.

Class Proxy

Unlike existing methods (Zhou et al. 2022c; Gao et al. 2024) that categorize the background class as a standard class, we define the background class as the complement of the target classes, thus more accurately reflecting its inherent nature. Consequently, we exclusively assign class proxies, which are randomly initialized, to the foreground classes. Formally, at session s, the set of class proxies can be represented as:

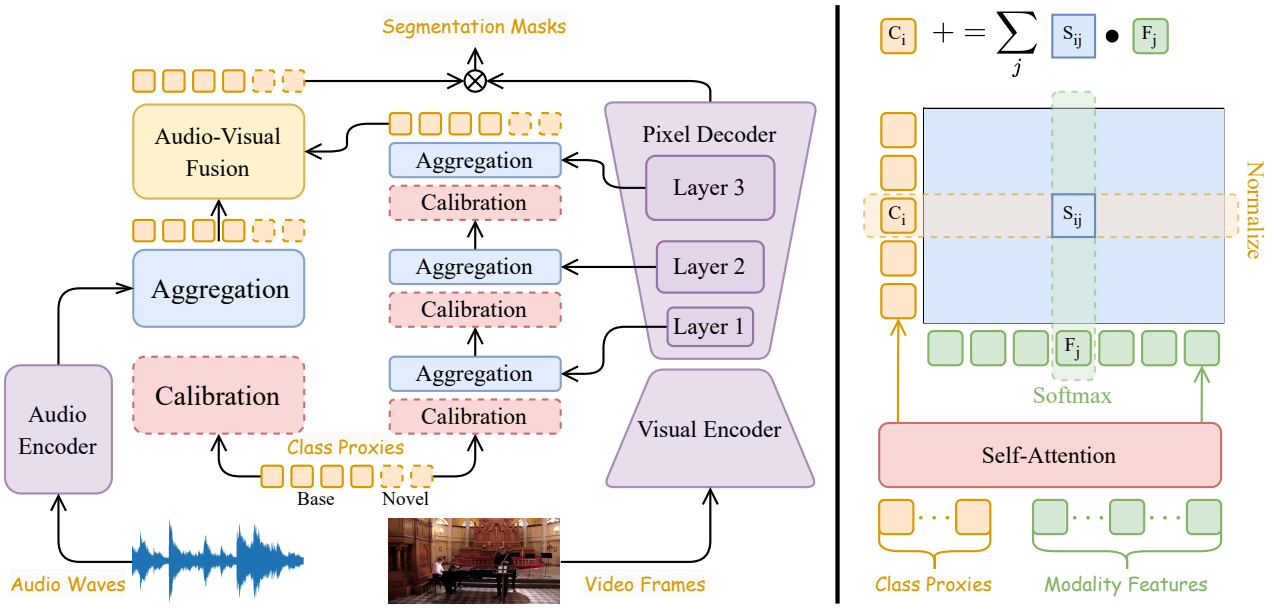
$$P = \{p^c \mid c \in C^0 \cup \cdots \cup C^s\}. \tag{2}$$

By constructing a one-vs-rest relationship for each class proxy (i.e., determining whether a pixel belongs to a specific class), we implicitly model the background class by considering the complement of all foreground classes.

Proxy Aggregation

For each foreground class, we leverage its proxy to independently aggregate class-specific features from the audio features F_a and video features F_v.

Proxy Feature Alignment. We first flatten the height and width dimensions of the visual feature at each scale, and then the flattened visual features and audio features are respectively concatenated with the class proxies. Subsequently,



(a) Model Architecture (b) Proxy Aggregation Module

Figure 2: **(a) Overview.** Our model employs a class-centric dual-tower architecture. Each foreground class is explicitly assigned a cross-modal class proxy. In each modality tower, the proxy aggregation modules are employed to independently aggregate class-specific audio or visual features. Then, cross-modal proxy fusion is conducted on a class-wise basis at a later stage. The solid-lined components constitute our base model, while the dashed-lined ones are extensions for the few-shot incremental sessions. **(b) Module Details.** The proxy aggregation module initially performs self-attention on the concatenated class proxies and modality features to align them. Subsequently, it calculates the similarities between aligned class proxies and modality features and uses these measures as weights to selectively assign features to the corresponding class proxies.

self-attention layers (Vaswani et al. 2017) are applied to align audio and visual features with the class proxies:

$$X_a = [F_a; P], X_v^1 = [F_v^1; P]; \tag{3}$$

$$\begin{aligned}[\hat{F}_a, \hat{P}_a] &= \text{Softmax}(X_a W_a^q (X_a W_a^k)^\top) X_a W_a^v, \\ [\hat{F}_v^1; \hat{P}_v^1] &= \text{Softmax}(X_v^1 W_v^{q,1} (X_v^1 W_v^{k,1})^\top) X_v^1 W_v^{v,1}.\end{aligned} \tag{4}$$

Coarse-to-Fine Aggregation. Since we have multi-scale visual features, we additionally pass the coarse class proxies $\hat{F}_v^1$ obtained from the first aggregation layer of the visual branch to the subsequent aggregation layers for interaction with finer visual features. The coarse-to-fine aggregation process allows for the gradual refinement of the class proxies, which can be written as:

$$X_v^i = [F_v^i; \hat{P}_v^{i-1}], i \in [2, 3]; \tag{5}$$

$$[\hat{F}_v^i; \hat{P}_v^i] = \text{Softmax}(X_v^i W_v^{q,i} (X_v^i W_v^{k,i})^\top) X_v^i W_v^{v,i}, i \in [2, 3]. \tag{6}$$

In Equations (4) and (6), all of the matrices W_a^q, W_a^k, W_a^v and $W_v^{q,i}, W_v^{k,i}, W_v^{v,i}$ ($i \in [1, 2, 3]$) denote the learnable weights of the linear projections for features and class proxies of the audio and visual modalities. The notation $[\cdot;\cdot]$ represents the concatenation operator, and Softmax denotes the softmax operation along the last dimension.

Subsequently, we aggregate class-specific features from the modality-specific class proxies and modality features obtained from self-attention layers to update class proxies.

Proxy-Feature Similarity. we first compute the similarity between each class proxy and each corresponding modality feature (only one feature for audio and $h \times w$ features for each visual scale, where (h, w) denote the resolution of that scale), resulting in an audio similarity map and visual similarity maps. From the perspective of each feature, we apply a softmax operation to compute the probability of each feature belonging to the classes. From the viewpoint of each class proxy, to determine the contribution of each feature to the output proxy representation of that class, we normalize across all features. The similarity map can be calculated as:

$$\begin{aligned}S_a &= \text{Softmax}(\hat{P}_a \tilde{W}_a^q (\hat{F}_a \tilde{W}_a^k)^\top, -2), \\ S_v^i &= \text{Normalize}(\text{Softmax}(\hat{P}_v^i \tilde{W}_v^{q,i} (\hat{F}_v^i \tilde{W}_v^{k,i})^\top, -2), -1),\end{aligned} \tag{7}$$

where $\text{Softmax}(\cdot, axis)$ and $\text{Normalize}(\cdot, axis)$ represent the operations of applying softmax and normalization along the specified dimension $axis$, respectively; $i \in [1, 2, 3]$, $\tilde{W}_a^q, \tilde{W}_a^k$ as well as $\tilde{W}_v^{q,i}, \tilde{W}_v^{k,i}$ are another set of learnable weights for the linear projections.

Feature-to-Proxy Assignment. We assign the audio and visual features, weighted by the audio similarity map S_a and visual similarity maps $\{S_v^i\}_{i=1}^3$, to update the class proxy representations in a residual manner (He et al. 2016):

$$\begin{aligned}\bar{P}_a &= \hat{P}_a + S_a (\hat{F}_a \tilde{W}_a^v) \tilde{W}_a^o, \\ \bar{P}_v^i &= \hat{P}_v^i + S_v^i (\hat{F}_v^i \tilde{W}_v^{v,i}) \tilde{W}_v^{o,i},\end{aligned} \tag{8}$$

where $i \in [1, 2, 3]$, $\tilde{W}_a^v, \tilde{W}_a^o, \tilde{W}_v^{v,i}, \tilde{W}_v^{o,i}$ denote the learnable weights of linear projections for features and outputs in terms of audio and visual modalities.

Notably, our aggregation applies softmax over class proxies and normalizes over modality features, in contrast to applying softmax over features in Attention (Vaswani et al. 2017). This distinction fosters a competitive dynamic among proxies for feature assignment, which compels each proxy to aggregate class-specific features and thus better aligns with our class-centric design philosophy.

Proxy Calibration

The class proxies P in session s consist of two parts: base class proxies $P_b = \{p^c \mid c \in C^0\}$ and novel class proxies $P_n = \{p^c \mid c \in C^1 \cup \cdots \cup C^s\}$. The former is learned from abundant examples, while the latter is optimized with only few shots. To overcome the overfitting issue of novel proxies caused by data scarcity, we propose proxy calibration to transfer knowledge from base class proxies, enhancing the representation of novel class proxies based on the similarities between novel classes and base classes. Specifically, the proxy calibration module is implemented as the cross-attention (Vaswani et al. 2017), where novel proxies serve as query and base proxies act as key and value:

$$\tilde{P}_n = \mathrm{Softmax}(P_n W^q (P_b W^k)^\top) P_b W^v. \tag{9}$$

Here, W^q, W^k, W^v represent learnable projection matrices for the query, key, and value, respectively.

We position the proxy calibration module ahead of each proxy aggregation module, such that the novel proxies fed into the proxy aggregation module are already calibrated. Considering the distinct class similarities across different modalities, the calibration modules for each modality are isolated. In contrast, since class similarities at varying levels of granularity within the same modality are consistent, each modality tower shares the same calibration module.

Proxy Fusion

Through proxy aggregation, we obtain audio-specific class proxies $\bar{P}_a$ and visual-specific class proxies $\bar{P}_v = \bar{P}_v^3$. Since audio-visual semantic segmentation requires identifying sound-producing objects, it is crucial to align the features of audio and visual modalities before mask decoding.

For class c, its corresponding audio proxy $\bar{P}_a^c$ and visual proxy $\bar{P}_v^c$ are first processed through cross-attention, using the audio proxy as the query and the visual proxy as the key and value. The cross-attention produces an audio-visual class proxy P_{av}^c that is then element-wise multiplied with $\bar{P}_v^c$ to selectively amplify or suppress different visual channels under the guidance of the audio proxy $\bar{P}_a^c$, expressed as:

$$P_{av}^c = (\bar{P}_v^c W_v^v) \mathrm{Softmax}((\bar{P}_v^c W_v^k)^\top \bar{P}_a^c W_a, -2) \odot \bar{P}_v^c. \tag{10}$$

Here, W_a, W_v^k, and W_v^v denote the learnable weights associated with audio proxies and visual proxies.

The segmentation mask (prior to binarization) for class c, denoted as M^c, is generated by matrix multiplication of the fused class proxy P_{av}^c with the mask features F_m, followed by a sigmoid operation, which can be formalized as:

$$M^c = \mathrm{Sigmoid}(P_{av}^c F_m). \tag{11}$$

Training

Pipeline. The training pipeline comprises three sequential stages: the base training stage, the episode training stage, and the incremental training stage. (1) The base training stage uses abundant data from the base session to train the base model, and the proxy calibration modules are not incorporated at this stage. (2) The episode training stage constructs pseudo-incremental scenarios to train the proxy calibration module while freezing the base model. The pseudo-code can be found in Algorithm 1. (3) The incremental training stage focuses solely on learning the novel class proxies introduced in each incremental session from few shots, and the rest of the extended model remains fixed.

Algorithm 1: Episode training of proxy calibration module with pseudo-incremental scenarios. Let $\mathcal{N}$ denote the proxy calibration modules of both the audio and visual towers.

Input: Base classes C^0, base data D^0_{train}, base model $\mathcal{M}$.
Output: Proxy calibration module $\mathcal{N}$.

```
Randomly initialize N.
while not done do
    Sample pseudo-incremental classes I from C^0.
    Reset class proxies P_I corresponding to I in M.
    Sample support set S_I and query set Q_I from D^0_train.
    Optimize P_I with S_I and Equation (12).
    Optimize N with Q_I and Equation (12).
    Restore P_I to their original state in M.
end while
```

Loss Function. The loss function includes binary cross-entropy loss and Dice loss (Milletari, Navab, and Ahmadi 2016) for the masks corresponding to target classes C^s of the current session s, formally defined as follows:

$$\mathcal{L}(M, M_{gt}) = \mathcal{L}_{bce}(M, M_{gt}) + \lambda \mathcal{L}_{dice}(M, M_{gt}), \tag{12}$$

where $M = \{M^c \mid c \in C^s\}$ and M_{gt} denotes the masks of target classes output by the model and the Ground Truth, λ represents the weight of the Dice loss $\mathcal{L}_{dice}$ relative to the BCE loss $\mathcal{L}_{bce}$. Since non-target classes are labeled as background, we ignore those masks during training.

Inference

During the inference phase of session s, we combine the masks of all foreground classes to derive the final semantic mask. Given that the predicted mask for each class represents the probability of each pixel belonging to that class, the mask for the background class M^0 can be computed as:

$$M^0 = \max(1 - \sum_{c \in C^0 \cup \cdots \cup C^s} M^c, 0). \tag{13}$$

On the other hand, the mask can be interpreted as the model's confidence in assigning a pixel to a particular class. In this context, the probabilities of different masks at the same pixel position can be compared. Thus, the final semantic mask $\tilde{M}$ can be specified as:

$$\tilde{M} = \mathrm{argmax}([M^0, M]). \tag{14}$$

Experiments

Experimental Protocols

To evaluate the performance in the few-shot incremental setting, we construct the FSIL-AVSS benchmark using the AVSBench-Semantic dataset (Zhou et al. 2024), which contains 70 foreground classes and a background class. Of these, 50 are designated as base classes, while the remaining 20 are novel classes. The protocol begins with training on a large-scale dataset of base classes, followed by 4 few-shot incremental sessions, each introducing 5 novel classes.

The training set for the base session consists of all videos containing at least one sound-producing pixel from any base class and the paired audio segments, with labels retained only for the base classes and the rest set as background. During the incremental session, the support set comprises K-shot audio-visual pairs, where each shot consists of a video frame and the corresponding audio segment. To maintain semantic uniqueness, different shots are taken from different videos. Furthermore, each shot in the support set contains only one novel class, ensuring that each class appears exactly K times. Notably, each audio-visual pair in the training set across all sessions may contain sound-producing pixels from classes that will be learned in future sessions or have already been learned in past sessions. However, these pixels are labeled as background. We consider scenarios with 1 or 5 shots per class and average the results over multiple trials.

The evaluation set for each session incorporates all video frames with at least one sound-producing pixel from any previously learned class and the paired audio segments, while pixels from classes not yet learned are excluded from the metric calculation. Following the protocol of Cermelli et al., we assess the performance using three metrics based on the mean intersection-over-union (mIoU): mIoU on base classes (mIoU-B), mIoU on novel classes (mIoU-N), and their harmonic mean (mIoU-H). We report the evaluation results after the final incremental session.

In addition, we conduct experiments with the original AVSBench-Semantic dataset, treating it as a special case within our framework. Specifically, we consider all classes as the base classes and execute the base training stage. For all experiments on the AVSBench-Semantic benchmark, we strictly adhere to its initial setup and use mIoU and F-score as the evaluation metrics.

Implementation Details

We utilize the Pyramid Vision Transformer (PVT-v2) (Wang et al. 2022), pretrained on ImageNet (Russakovsky et al. 2015), as the visual encoder, and the VGGish (Hershey et al. 2017), pretrained on AudioSet (Gemmeke et al. 2017), as the audio encoder. The multi-scale deformable attention Transformer (Zhu et al. 2021) is employed as the pixel decoder. All video frames are resized to 224 $\times$ 224 pixels, and the audio signals are segmented into one-second clips. In the proxy aggregation module, the number of self-attention layers is set at 3 for the audio tower and 5 for the visual tower, respectively. We use the AdamW optimizer with a batch size of 4, an initial learning rate of 2 $\times$ 10-5, and train for 30 epochs. The hyperparameter λ in Equation (12) is set to 0.1.

Method	mIoU	F-score
3DC (Mahadevan et al. 2020)	17.3	21.6
AOT (Yang, Wei, and Yang 2021)	25.4	31.0
AVSBench (Zhou et al. 2024)	29.8	35.2
CATR (Li et al. 2023a)	32.8	38.5
BAVS (Liu et al. 2024a)	33.6	37.5
AVSegFormer (Gao et al. 2024)	36.7	42.0
Ours	**47.1**	**54.2**

Table 1: Comparison of several methods on the AVSBench-Semantic dataset with mIoU and F-score metrics.

Performance Comparison

We first conduct a comprehensive comparison between our model and existing methods (Mahadevan et al. 2020; Yang, Wei, and Yang 2021; Zhou et al. 2024; Li et al. 2023a; Liu et al. 2024a; Gao et al. 2024) on the AVSBench-Semantic benchmark. As shown in Table 1, our model substantially outperforms other methods, surpassing the previous leader by margins of 10.4 mIoU and 12.2 F-score, respectively.

We continue our experiments on the proposed few-shot incremental benchmark, comparing our model with the leading AVSegFormer, as well as its two adaptations implementing MiB (Cermelli et al. 2020) and PIFS (Cermelli et al. 2021). As shown in Table 2, AVSegFormer faces challenges with catastrophic forgetting and struggles to rapidly learn. Although MiB and PIFS moderately improve their performance, they are constrained by AVSegFormer and fail to address the unique challenges of AVSS. In contrast, our model significantly surpasses all baselines across all metrics.

Ablation Studies

Impact of Proxy Aggregation. To demonstrate the effectiveness of our aggregation module, we conduct experiments where it is replaced with a conventional Multihead Attention module (Vaswani et al. 2017). Specifically, the attention module includes self-attention among class proxies, cross-attention between class proxies and modality features, along with two layers of feedforward MLP. As illustrated in Table 3, models utilizing our aggregation module consistently outperform those employing the Attention module.

Impact of Dual Tower. We compare the dual-tower and single-tower architectures by adapting our model to a single-tower variant. In this setup, we move the cross-modal fusion module forward to fuse audio and visual features at each scale, resulting in multi-scale audio-visual features which are then fed into the visual tower, with the audio tower discarded. As illustrated in Table 3, the dual-tower models consistently surpass the single-tower models.

Impact of Proxy Calibration. To validate the indispensability of the proxy calibration modules, we conduct experiments by removing them in the 5-shot incremental scenario. As shown in Table 4, the mIoU-N drops significantly in their absence, primarily due to overfitting. Moreover, we omit the episode training stage and allow the randomly initialized proxy calibration modules to participate in training during each incremental session, which yields worse results.

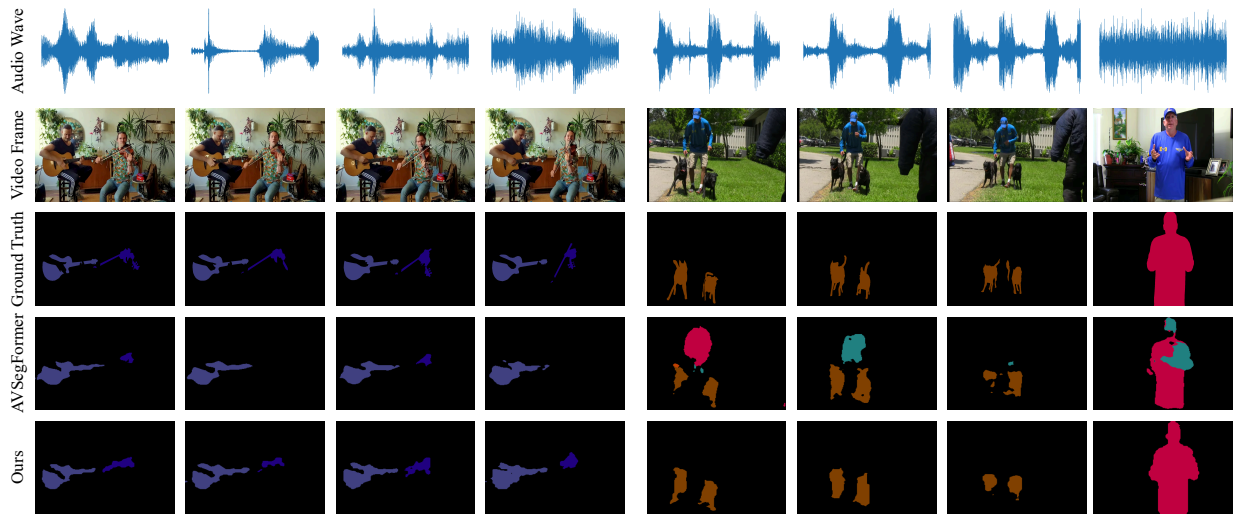


Figure 3: Visualization of cases with multiple sound sources across our model and the previously leading model.

Method	1-shot Incremental			5-shot Incremental		
	mIoU-B	mIoU-N	mIoU-H	mIoU-B	mIoU-N	mIoU-H
AVSegFormer (Gao et al. 2024)	10.4	0.1	0.2	5.0	3.8	4.3
AVSegFormer+MiB (Cermelli et al. 2020)	25.2	0.9	1.7	29.2	7.0	11.3
AVSegFormer+PIFS (Cermelli et al. 2021)	30.3	1.7	3.2	34.0	11.4	17.1
Ours	**44.9**	**5.9**	**10.4**	**43.9**	**19.2**	**26.7**

Table 2: Comparison of different methods across 1-shot and 5-shot incremental scenarios with three mIoU metrics.

Model Architecture	Module Function	Standard		5-shot Incremental		
		mIoU	F-score	mIoU-B	mIoU-N	mIoU-H
Single Tower	Attention (Vaswani et al. 2017)	35.7	41.6	34.1	4.3	7.6
Dual Tower	Attention (Vaswani et al. 2017)	37.0	42.8	34.7	10.9	16.6
Single Tower	Proxy Aggregation	38.5	44.4	36.3	7.9	13.0
Dual Tower	Proxy Aggregation	47.1	54.2	43.9	19.2	26.7

Table 3: Ablation studies of various model architectures and module functions across standard and 5-shot incremental scenarios.

Calibration	Episode	mIoU-B	mIoU-N	mIoU-H
		36.6	2.7	5.0
✓		38.4	14.8	21.4
✓	✓	43.9	19.2	26.7

Table 4: Ablation studies of the proxy calibration module in the 5-shot incremental scenario.

Qualitative Analysis. In Figure 3, we visualize several multi-source cases. In the left case, when one source dominates, AVSegFormer misses the subordinate sources (frames 2 and 4), whereas ours remains unaffected and consistently segments all sound producers. In the right case, AVSegFormer mistakenly recognizes a person as producing sound (frames 1 and 2) and also misclassifies objects (frames 2 and 4), while ours maintains accurate identification throughout.

Conclusion

In conclusion, the FINGER model represents a significant advancement in the field of audio-visual semantic segmentation (AVSS) through its pioneering class-centric and dual-tower framework. It successfully tackles the intricate challenges of few-shot incremental learning in audio-visual semantic segmentation (FSIL-AVSS), including foreground modeling and knowledge inheritance. The class-centric foreground aggregation design ensures that the background annotations do not interfere with foreground modeling, while the dual-tower knowledge transfer mechanism enables effective cross-class knowledge transfer when learning from limited incremental samples. Extensive experiments demonstrate that FINGER outperforms state-of-the-art methods on both the AVSS and FSIL-AVSS benchmarks, highlighting its potential to drive future innovations in the field.

Acknowledgments

This work is supported by the Advanced Research and Technology Innovation Centre (ARTIC), the National University of Singapore under Grant (project number: ELDT-RP2), and the RIE2025 Career Development Fund (Award C233312009), administered by A*STAR.

References

Akyürek, A. F.; Akyürek, E.; Wijaya, D. T.; and Andreas, J. 2022. Subspace regularizers for few-shot class incremental learning. In *ICLR*.

Arandjelovic, R.; and Zisserman, A. 2017. Look, listen and learn. In *ICCV*, 609–617.

Arandjelovic, R.; and Zisserman, A. 2018. Objects that sound. In *ECCV*, 435–451.

Cermelli, F.; Mancini, M.; Bulo, S. R.; Ricci, E.; and Caputo, B. 2020. Modeling the background for incremental learning in semantic segmentation. In *CVPR*, 9233–9242.

Cermelli, F.; Mancini, M.; Xian, Y.; Akata, Z.; and Caputo, B. 2021. Prototype-based incremental few-shot semantic segmentation. In *BMVC*.

Chen, H.; Xie, W.; Afouras, T.; Nagrani, A.; Vedaldi, A.; and Zisserman, A. 2021. Localizing visual sounds the hard way. In *CVPR*, 16867–16876.

Cheraghian, A.; Rahman, S.; Fang, P.; Roy, S. K.; Petersson, L.; and Harandi, M. 2021. Semantic-aware knowledge distillation for few-shot class-incremental learning. In *CVPR*, 2534–2543.

Chi, Z.; Gu, L.; Liu, H.; Wang, Y.; Yu, Y.; and Tang, J. 2022. Metafscil: A meta-learning approach for few-shot class incremental learning. In *CVPR*, 14166–14175.

Dong, S.; Hong, X.; Tao, X.; Chang, X.; Wei, X.; and Gong, Y. 2021. Few-shot class-incremental learning via relation knowledge distillation. In *AAAI*, 1255–1263.

Fang, X.; Easwaran, A.; Genest, B.; and Suganthan, P. N. 2024a. Your Data Is Not Perfect: Towards Cross-Domain Out-of-Distribution Detection in Class-Imbalanced Data. *ESWA*.

Fang, X.; Fang, W.; Liu, D.; Qu, X.; Dong, J.; Zhou, P.; Li, R.; Xu, Z.; Chen, L.; Zheng, P.; et al. 2024b. Not all inputs are valid: Towards open-set video moment retrieval using language. In *ACM MM*.

Fang, X.; Fang, W.; Liu, D.; and Zhou, P. 2025. Multi-Pair Temporal Sentence Grounding via Multi-Thread Knowledge Transfer Network. In *AAAI*.

Fang, X.; Liu, D.; Fang, W.; Zhou, P.; Xu, Z.; Xu, W.; Chen, J.; and Li, R. 2024c. Fewer Steps, Better Performance: Efficient Cross-Modal Clip Trimming for Video Moment Retrieval Using Language. In *AAAI*.

Fang, X.; Liu, D.; Zhou, P.; and Nan, G. 2023. You can ground earlier than see: An effective and efficient pipeline for temporal sentence grounding in compressed videos. In *CVPR*.

Gao, S.; Chen, Z.; Chen, G.; Wang, W.; and Lu, T. 2024. Avsegformer: Audio-visual segmentation with transformer. In *AAAI*, 12155–12163.

Gemmeke, J. F.; Ellis, D. P.; Freedman, D.; Jansen, A.; Lawrence, W.; Moore, R. C.; Plakal, M.; and Ritter, M. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *ICASSP*, 776–780.

Hao, D.; Mao, Y.; He, B.; Han, X.; Dai, Y.; and Zhong, Y. 2024. Improving audio-visual segmentation with bidirectional generation. In *AAAI*, 2067–2075.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

Hersche, M.; Karunaratne, G.; Cherubini, G.; Benini, L.; Sebastian, A.; and Rahimi, A. 2022. Constrained few-shot class-incremental learning. In *CVPR*, 9057–9067.

Hershey, S.; Chaudhuri, S.; Ellis, D. P.; Gemmeke, J. F.; Jansen, A.; Moore, R. C.; Plakal, M.; Platt, D.; Saurous, R. A.; Seybold, B.; et al. 2017. CNN architectures for large-scale audio classification. In *ICASSP*, 131–135.

Ji, W.; Li, L.; Fei, H.; Liu, X.; Yang, X.; Li, J.; and Zimmermann, R. 2023a. Towards complex-query referring image segmentation: A novel benchmark. *ACM TOMM*.

Ji, W.; Liang, R.; Liao, L.; Fei, H.; and Feng, F. 2023b. Partial annotation-based video moment retrieval via iterative learning. In *ACM MM*, 4330–4339.

Ji, W.; Liang, R.; Zheng, Z.; Zhang, W.; Zhang, S.; Li, J.; Li, M.; and Chua, T.-s. 2023c. Are binary annotations sufficient? video moment retrieval via hierarchical uncertainty-based active learning. In *CVPR*, 23013–23022.

Ji, W.; Qin, Y.; Chen, L.; Wei, Y.; Wu, Y.; and Zimmermann, R. 2024a. Mrtnet: Multi-resolution temporal network for video sentence grounding. In *ICASSP*, 2770–2774. IEEE.

Ji, W.; Shi, R.; Wei, Y.; Zhao, S.; and Zimmermann, R. 2024b. Weakly Supervised Video Moment Retrieval via Location-irrelevant Proposal Learning. In *TheWebConf*, 1595–1603.

Kang, H.; Yoon, J.; Madjid, S. R. H.; Hwang, S. J.; and Yoo, C. D. 2023. On the soft-subnetwork for few-shot class incremental learning. In *ICLR*.

Kim, D.-Y.; Han, D.-J.; Seo, J.; and Moon, J. 2023. Warping the space: Weight space rotation for class-incremental few-shot learning. In *ICLR*.

Li, K.; Yang, Z.; Chen, L.; Yang, Y.; and Xiao, J. 2023a. Catr: Combinatorial-dependence audio-queried transformer for audio-visual video segmentation. In *ACM MM*, 1485–1494.

Li, M.; Wang, H.; Zhang, W.; Miao, J.; Zhao, Z.; Zhang, S.; Ji, W.; and Wu, F. 2023b. Winner: Weakly-supervised hierarchical decomposition and alignment for spatio-temporal video grounding. In *CVPR*, 23090–23099.

Li, M.; Wang, T.; Zhang, H.; Zhang, S.; Zhao, Z.; Miao, J.; Zhang, W.; Tan, W.; Wang, J.; Wang, P.; et al. 2022. End-to-End Modeling via Information Tree for One-Shot Natural Language Spatial Video Grounding. In *ACL*, 8707–8717.

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017. Feature pyramid networks for object detection. In *CVPR*, 2117–2125.

Liu, C.; Li, P.; Zhang, H.; Li, L.; Huang, Z.; Wang, D.; and Yu, X. 2024a. BAVS: bootstrapping audio-visual segmentation by integrating foundation knowledge. *TMM*.

Liu, J.; Liu, Y.; Zhang, F.; Ju, C.; Zhang, Y.; and Wang, Y. 2024b. Audio-Visual Segmentation via Unlabeled Frame Exploitation. In *CVPR*, 26328–26339.

Mahadevan, S.; Athar, A.; Ošep, A.; Hennen, S.; Leal-Taixé, L.; and Leibe, B. 2020. Making a case for 3d convolutions for object segmentation in videos. In *BMVC*.

Mao, Y.; Zhang, J.; Xiang, M.; Zhong, Y.; and Dai, Y. 2023. Multimodal variational auto-encoder based audio-visual segmentation. In *ICCV*, 954–965.

Milletari, F.; Navab, N.; and Ahmadi, S.-A. 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *3DV*, 565–571.

Mo, S.; and Morgado, P. 2022. Localizing visual sounds the easy way. In *ECCV*, 218–234.

Mo, S.; and Tian, Y. 2022. Multi-modal grouping network for weakly-supervised audio-visual video parsing. *NeurIPS*, 35: 34722–34733.

Mo, S.; and Tian, Y. 2023. Av-sam: Segment anything model meets audio-visual localization and segmentation. *arXiv preprint arXiv:2305.01836*.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *IJCV*, 115: 211–252.

Shan, L.; Zhou, W.; and Zhao, G. 2023. Incremental few shot semantic segmentation via class-agnostic mask proposal and language-driven classifier. In *ACM MM*, 8561–8570.

Shi, G.; Wu, Y.; Liu, J.; Wan, S.; Wang, W.; and Lu, T. 2022. Incremental few-shot semantic segmentation via embedding adaptive-update and hyper-class representation. In *ACM MM*, 5547–5556.

Song, Z.; Zhao, Y.; Shi, Y.; Peng, P.; Yuan, L.; and Tian, Y. 2023. Learning with fantasy: Semantic-aware virtual contrastive constraint for few-shot class-incremental learning. In *CVPR*, 24183–24192.

Tao, X.; Hong, X.; Chang, X.; Dong, S.; Wei, X.; and Gong, Y. 2020. Few-shot class-incremental learning. In *CVPR*, 12183–12192.

Tian, Y.; Li, D.; and Xu, C. 2020. Unified multisensory perception: Weakly-supervised audio-visual video parsing. In *ECCV*, 436–454.

Tian, Y.; Shi, J.; Li, B.; Duan, Z.; and Xu, C. 2018. Audio-visual event localization in unconstrained videos. In *ECCV*, 247–263.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is all you need. *NeurIPS*.

Vinyals, O.; Blundell, C.; Lillicrap, T.; Wierstra, D.; et al. 2016. Matching networks for one shot learning. *NeurIPS*, 29.

Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; and Shao, L. 2022. Pvt v2: Improved baselines with pyramid vision transformer. *CVMJ*, 8(3): 415–424.

Wang, Y.; Liu, W.; Li, G.; Ding, J.; Hu, D.; and Li, X. 2024. Prompting segmentation with sound is generalizable audio-visual source localizer. In *AAAI*, 5669–5677.

Wu, Y.; and Yang, Y. 2021. Exploring heterogeneous clues for weakly-supervised audio-visual video parsing. In *CVPR*, 1326–1335.

Wu, Y.; Zhu, L.; Yan, Y.; and Yang, Y. 2019. Dual attention matching for audio-visual event localization. In *ICCV*, 6292–6300.

Xiu, J.; Li, M.; Ji, W.; Chen, J.; Zhao, H.; Satoh, S.; and Zimmermann, R. 2024. Hierarchical Debiasing and Noisy Correction for Cross-domain Video Tube Retrieval. In *ACM MM*, 9271–9280.

Yang, B.; Lin, M.; Zhang, Y.; Liu, B.; Liang, X.; Ji, R.; and Ye, Q. 2022. Dynamic support network for few-shot class incremental learning. *TPAMI*, 45(3): 2945–2951.

Yang, Q.; Nie, X.; Li, T.; Gao, P.; Guo, Y.; Zhen, C.; Yan, P.; and Xiang, S. 2024. Cooperation Does Matter: Exploring Multi-Order Bilateral Relations for Audio-Visual Segmentation. In *CVPR*, 27134–27143.

Yang, Y.; Yuan, H.; Li, X.; Lin, Z.; Torr, P.; and Tao, D. 2023. Neural collapse inspired feature-classifier alignment for few-shot class incremental learning. In *ICLR*.

Yang, Z.; Wei, Y.; and Yang, Y. 2021. Associating objects with transformers for video object segmentation. *NeurIPS*, 34: 2491–2502.

Zhang, C.; Song, N.; Lin, G.; Zheng, Y.; Pan, P.; and Xu, Y. 2021. Few-shot incremental learning with continually evolved classifiers. In *CVPR*, 12455–12464.

Zhou, D.-W.; Wang, F.-Y.; Ye, H.-J.; Ma, L.; Pu, S.; and Zhan, D.-C. 2022a. Forward compatible few-shot class-incremental learning. In *CVPR*, 9046–9056.

Zhou, D.-W.; Ye, H.-J.; Ma, L.; Xie, D.; Pu, S.; and Zhan, D.-C. 2022b. Few-shot class-incremental learning by sampling multi-phase tasks. *TPAMI*, 45(11): 12816–12831.

Zhou, J.; Shen, X.; Wang, J.; Zhang, J.; Sun, W.; Zhang, J.; Birchfield, S.; Guo, D.; Kong, L.; Wang, M.; et al. 2024. Audio-visual segmentation with semantics. *IJCV*, 1–21.

Zhou, J.; Wang, J.; Zhang, J.; Sun, W.; Zhang, J.; Birchfield, S.; Guo, D.; Kong, L.; Wang, M.; and Zhong, Y. 2022c. Audio–visual segmentation. In *ECCV*, 386–403.

Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2021. Deformable detr: Deformable transformers for end-to-end object detection. In *ICLR*.