AVSegFormer: Audio-Visual Segmentation with Transformer

Shengyi Gao[1], Zhe Chen[1], Guo Chen[1], Wenhai Wang[2], Tong Lu[1]*

[1]State Key Lab for Novel Software Technology, Nanjing University
[2]The Chinese University of Hong Kong
lutong@nju.edu.cn

Abstract

Audio-visual segmentation (AVS) aims to locate and segment the sounding objects in a given video, which demands audio-driven pixel-level scene understanding. The existing methods cannot fully process the fine-grained correlations between audio and visual cues across various situations dynamically. They also face challenges in adapting to complex scenarios, such as evolving audio, the coexistence of multiple objects, and more. In this paper, we propose AVSegFormer, a novel framework for AVS that leverages the transformer architecture. Specifically, It comprises a dense audio-visual mixer, which can dynamically adjust interested visual features, and a sparse audio-visual decoder, which implicitly separates audio sources and automatically matches optimal visual features. Combining both components provides a more robust bidirectional conditional multi-modal representation, improving the segmentation performance in different scenarios. Extensive experiments demonstrate that AVSegFormer achieves state-of-the-art results on the AVS benchmark. The code is available at https://github.com/vvvb-github/AVSegFormer.

Introduction

Just as humans effortlessly establish meaningful connections between audio and visual signals, capturing the rich information they convey, the intertwined modalities of audio and vision play pivotal roles in observing and comprehending the real world. Based on this insight, a wide range of audio-visual understanding tasks, such as audio-visual correspondence (Arandjelovic and Zisserman 2017, 2018), audio-visual event localization (Lin, Li, and Wang 2019; Lin and Wang 2020), audio-visual video parsing (Tian, Li, and Xu 2020; Wu and Yang 2021), and sound source localization (Arandjelovic and Zisserman 2017, 2018) have been proposed and actively explored in recent research.

Unlike these coarse-grained tasks, audio-visual segmentation (AVS) (Zhou et al. 2023) proposes more fine-grained perceptive goals, aiming to locate the audible frames and delineate the shape of the sounding objects (Zhou et al. 2022, 2023). To be more specific, this task involves three sub-tasks: single sound source segmentation (S4), multiple sound source segmentation (MS3), and audio-visual semantic segmentation (AVSS). Figure 1 illustrates the objectives of the three sub-tasks. Their fine-grained perceptual goal demands the model to possess the capability to discern the intricate relationship between each image pixel and audio information. However, the existing methods (Chen et al. 2021; Qian et al. 2020; Mahadevan et al. 2020) developed for other audio-visual tasks face challenges when directly applied in this context.

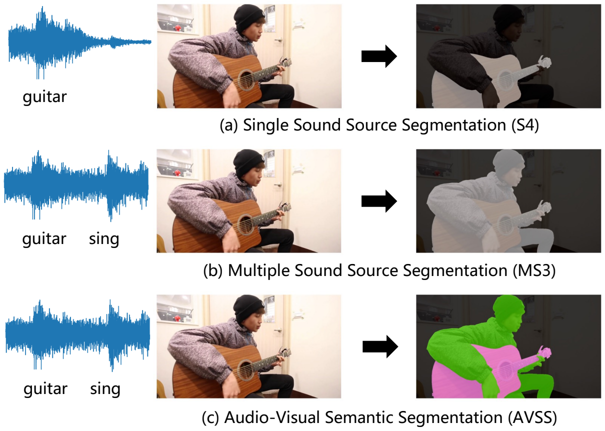
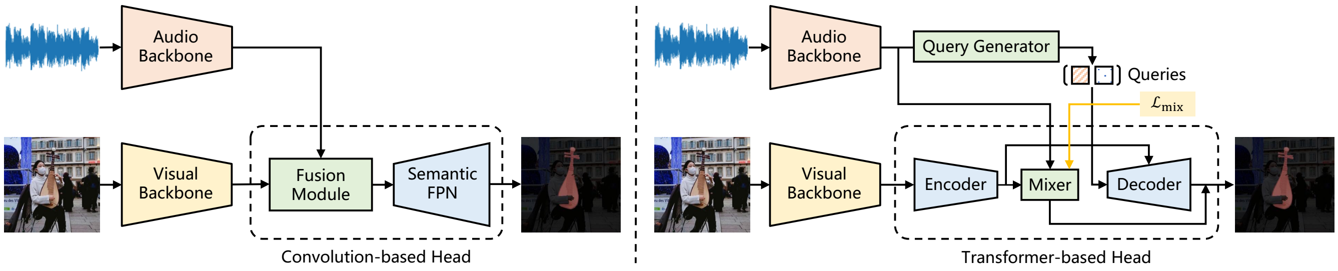

guitar
(a) Single Sound Source Segmentation (S4)
guitar
sing
(b) Multiple Sound Source Segmentation (MS3)
guitar
sing
(c) Audio-Visual Semantic Segmentation (AVSS)

Figure 1: Illustration of audio-visual segmentation (AVS). AVS aims to segment sounding objects from video frames according to the given audio. In the S4 sub-task, the input audio only contains one sound source, while in MS3 the input audio has multiple sound sources. Besides, S4 and MS3 only require binary segmentation, whereas AVSS requires more difficult multiple-category semantic segmentation.

AVSBench (Zhou et al. 2023) first proposes the fine-grained perceptive method for AVS tasks that achieves state-of-the-art audio-visual segmentation performance. Figure 2(a) illustrates its network architecture, which incorporates a modality fusion module before Semantic FPN (Kirillov et al. 2019) to enable audio-visual segmentation. This method is simple and effective yet falls short of fully mining the fine-grained correlations between audio and visual cues across various situations. First, a series of targets sound simultaneously in scenes of multiple sound sources. The mixed audio signal with higher information density is difficult to attend to the visual signal adaptively. Second, the simple-fusion model may make it difficult to extract au-

[*]corresponding author



(a) AVSBench Baseline

(b) AVSegFormer (ours)

Figure 2: Overview of the AVSBench baseline and our AVSegFormer. (a) The baseline method (Zhou et al. 2022) incorporates a modality fusion module before Semantic FPN (Li et al. 2022b). (b) The proposed AVSegFormer performs audio-visual segmentation with transformer-based architecture. It has four key designs to significantly improve performance, including (1) a transformer encoder, (2) an audio-visual mixer, (3) a query generator, and (4) a cross-modal transformer decoder.

dio information bound with the corresponding frame when the sound source objects change over time. Additionally, in cases where multiple objects coexist within a single contextual frame (*e.g.*, two guitars or one person and one guitar), the vanilla dense segmentation method struggles to untangle the different sound sources in the audio, making it challenging to achieve precise localization and segmentation in a one-to-one manner.

To remedy these issues, we propose **AVSegFormer**, a novel framework for audio-visual segmentation with the transformer architecture. The brief architecture is shown in Figure 2(b). AVSegFormer comprises four key components: (1) a transformer encoder building the mask feature, (2) an audio-visual mixer generating the vision-conditioned mask feature, (3) a query generator initializing sparse audio-conditioned queries, and (4) a cross-modal transformer decoder separating potential sparse object in the visual feature. Among them, we design an auxiliary mixing loss to supervise the cross-modal feature generated by the audio-visual mixer. It encourages the model to attend to useful information within complex audio semantics and predict the segmentation mask densely. Compared to the dense vision-conditioned mixers, the cross-modal transformer decoder aims to build potential sparse queries with the reversed condition. It implicitly separates audio sources and automatically matches optimal visual features for different queries. These queries will be combined with the vision-conditioned feature map through the matrix production. At last, the fused bidirectional conditional feature maps will be used for the final segmentation.

Overall, our contributions to this work are three-fold:

(1) We propose AVSegFormer, a transformer-based method for three scenarios of the audio-visual segmentation tasks. It combines bidirectional conditional cross-modal feature fusion to provide a more robust audio-visual segmentation representation.

(2) We propose a dense audio-visual mixer and a sparse audio-visual decoder to provide pixel-level and instance-level complementary representations that can efficiently adapt the scenarios of multiple sound sources and objects.

(3) Extensive experiments on three sub-tasks of AVS are conducted, demonstrating that AVSegFormer significantly outperforms existing state-of-the-art methods (Mao et al. 2023; Zhou et al. 2022).

Related Works

Multi-Modal Tasks

In recent years, multi-modal tasks have gained significant attention in the research community. Among these, text-visual tasks have attracted considerable interest from researchers. Numerous works focus on related tasks, such as visual question answering (Antol et al. 2015; Wu et al. 2016) and visual grounding (Deng et al. 2021; Kamath et al. 2021). In addition to text-visual tasks, audio-visual tasks are emerging as hot spots. Related tasks include audio-visual correspondence (Arandjelovic and Zisserman 2017, 2018), audio-visual event localization (Lin, Li, and Wang 2019; Lin and Wang 2020), and sound source localization (Arandjelovic and Zisserman 2017, 2018). Concurrently, many works (Zhu et al. 2022; Wang et al. 2022a) have proposed unified architectures to deal with multi-modal inputs.

Most of these works are based on transformer architecture (Vaswani et al. 2017), demonstrating a strong cross-modal capability. Their success highlights the reliability of transformers in the multi-modal field. As a recently proposed multi-modal task, audio-visual segmentation (Zhou et al. 2022, 2023) shares many commonalities with the aforementioned tasks. The pioneering works in these areas have significantly inspired our research of AVSegFormer.

Vision Transformer

During the past few years, Transformer (Vaswani et al. 2017) has experienced rapid development in natural language processing. Following this success, the Vision Transformer (ViT) (Dosovitskiy et al. 2020) emerged, bringing the transformer into the realm of computer vision and yielding impressive results. Numerous works (Liu et al. 2021; Wang et al. 2022b; Chen et al. 2022) have built upon ViT, leading to the maturation of vision transformers, especially in object detection and image segmentation tasks. As the performance of vision transformers continues to advance, they are

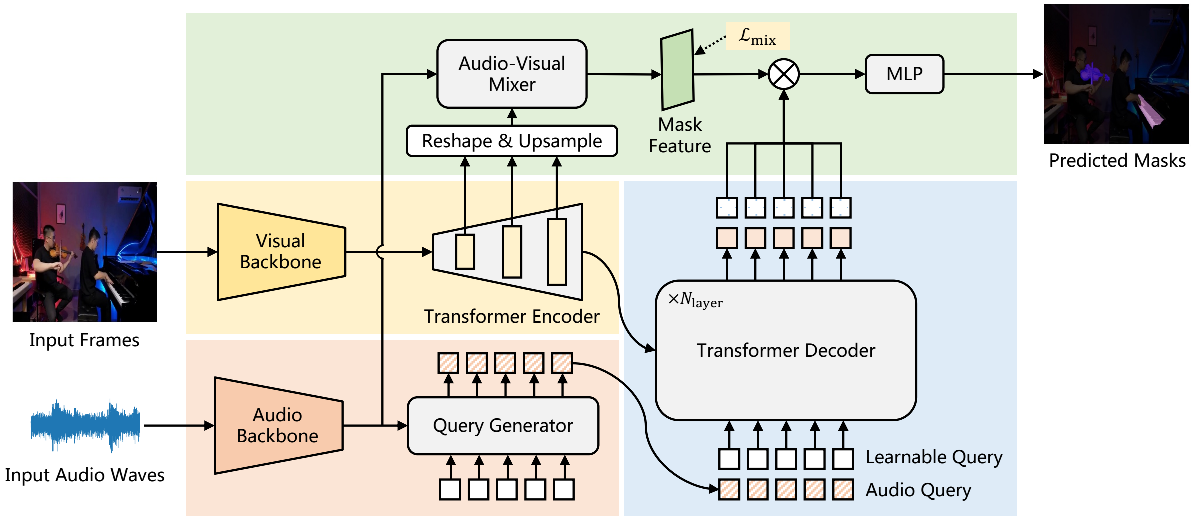


Figure 3: Overall architecture of AVSegFormer. We propose four key components in this framework: (1) The transformer encoder builds the initial mask feature; (2) An audio-visual mixer with an auxiliary mixing loss $\mathcal{L}_{\text{mix}}$ generates the vision-conditioned mask feature. (3) The query generator initializes sparse audio-conditioned queries, enabling the model to recognize abundant auditory semantics. (4) The cross-modal transformer decoder separates potential sparse objects in the visual feature.

increasingly replacing CNNs as the mainstream paradigm in the field of computer vision, especially in object detection and image segmentation tasks.

For downstream tasks, Carion et al. (2020) proposed the DETR model and designed a novel bipartite matching loss based on the transformer architecture. Subsequently, improved frameworks such as Deformable DETR (Zhu et al. 2020) and DINO (Zhang et al. 2022) are proposed, introducing mechanisms like deformable attention and denoise training. These arts take vision transformers to new heights. The remarkable performance of vision transformers has also inspired us to apply this paradigm to AVS tasks, anticipating further advancements in the field.

Image Segmentation

Image segmentation is a critical visual task that involves partitioning an image into distinct segments or regions. It includes three different tasks: instance segmentation, semantic segmentation, and panoptic segmentation. Early research proposed specialized models for these tasks, such as Mask R-CNN (He et al. 2017) and HTC (Chen et al. 2019) for instance segmentation, or FCN (Long, Shelhamer, and Darrell 2015) and U-Net (Ronneberger, Fischer, and Brox 2015) for semantic segmentation. After panoptic segmentation was proposed, some related research (Kirillov et al. 2019; Xiong et al. 2019; Li et al. 2022b) were conducted and designed universal models for both tasks.

The recent introduction of the transformer has led to the development of new models that can unify all the segmentation tasks. Mask2Former (Cheng et al. 2022) is one such model that introduces mask attention into the transformer. Mask DINO (Li et al. 2022a) is a unified transformer-based framework for both detection and segmentation. Recently, OneFormer (Jain et al. 2022) presented a new universal image segmentation framework with transformers. These models have brought image segmentation to a new level. Considering that the AVS task involves segmentation, these methods have significantly contributed to our work.

Methods

Overall Architecture

Figure 3 illustrates the overall architecture of our method. In contrast to previous CNN-based methods (Zhou et al. 2022, 2023), we design a query-based framework to leverage the transformer architecture. Specifically, the query generator initializes audio queries, and the transformer encoder extracts multi-scale features, which serve as inputs of the transformer decoder for separating potential sparse objects. Besides, the audio-visual mixer will further amplify relevant features and suppress irrelevant ones, while the auxiliary mixing loss helps supervise the reinforced features.

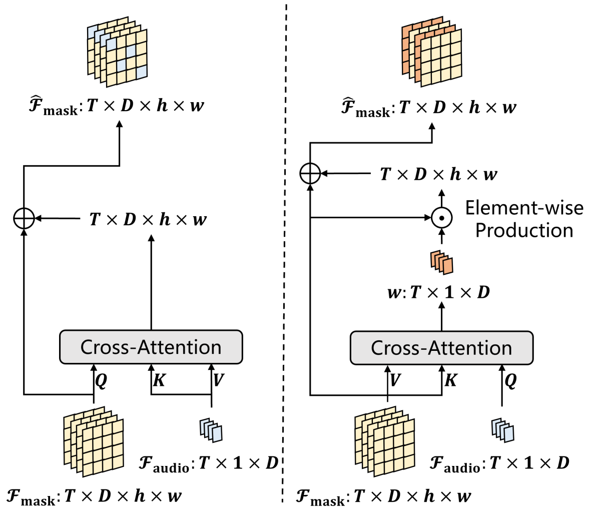
Multi-Modal Representation

Visual encoder. We follow the feature extraction process adopted in previous methods (Zhou et al. 2022, 2023), which uses a visual backbone and an audio backbone to extract video and audio features, respectively. The dataset provides pre-extracted frame images from videos, making the process similar to image feature extraction. Specifically, the input video frames are denoted as $x_{\text{visual}} \in \mathbb{R}^{T\times 3\times H\times W}$, in which T denotes the number of frames. Then, we use a visual backbone (*e.g.*, ResNet-50 (He et al. 2016)) to extract hierarchical visual features $\mathcal{F}_{\text{visual}}$, which can be written as:

$$\mathcal{F}_{\text{visual}} = \{\mathcal{F}_1, \mathcal{F}_2, \mathcal{F}_3, \mathcal{F}_4\}, \quad (1)$$

in which $\mathcal{F}_i \in \mathbb{R}^{T\times 256\times \frac{H}{2^{i+1}}\times \frac{W}{2^{i+1}}}$ and $i \in [1, 2, 3, 4]$.

Audio encoder. The process of audio feature extraction follows the VGGish (Hershey et al. 2017) method. Initially,


$\hat{\mathcal{F}}_{\text{mask}}$: $T \times D \times h \times w$
$T \times D \times h \times w$
Cross-Attention
Q
K
V
$\mathcal{F}_{\text{audio}}$: $T \times 1 \times D$
$\mathcal{F}_{\text{mask}}$: $T \times D \times h \times w$
$\hat{\mathcal{F}}_{\text{mask}}$: $T \times D \times h \times w$
$T \times D \times h \times w$
Element-wise Production
w: $T \times 1 \times D$
Cross-Attention
V
K
Q
$\mathcal{F}_{\text{audio}}$: $T \times 1 \times D$
$\mathcal{F}_{\text{mask}}$: $T \times D \times h \times w$

(a) Cross-Attention Mixer (CRA) (b) Channel-Attention Mixer (CHA)

Figure 4: Architecture of the audio-visual mixer. (a) Our initial design incorporates a cross-attention mixer, which fails to deliver satisfactory results. (b) We ultimately adopted the channel-attention mixer design, demonstrating significantly improved performance.

the audio is resampled to 16kHz mono audio $x_{\text{audio}} \in \mathbb{R}^{N_{\text{samples}} \times 96 \times 64}$, where N_{samples} is related to the audio duration. Then, we perform a short-time Fourier transform to obtain a mel spectrum. The mel spectrum is calculated by mapping the spectrum to a 64th-order mel filter bank and then fed into the VGGish model to obtain the audio features $\mathcal{F}_{\text{audio}} \in \mathbb{R}^{T \times 256}$, where T means the number of frames.

Query Generator

The query generator is designed to generate sparse audio-conditioned queries, which helps the model fully understand the auditory information. At the beginning, we have an initial query $Q_{\text{init}} \in \mathbb{R}^{T \times N_{\text{query}} \times D}$ and audio feature $\mathcal{F}_{\text{audio}} \in \mathbb{R}^{T \times D}$, here N_{query} represents the number of queries. We employ Q_{init} as queries and $\mathcal{F}_{\text{audio}}$ as keys and values, feed them into the query generator, and obtain the audio query $Q_{\text{audio}} \in \mathbb{R}^{T \times N_{\text{query}} \times D}$. Finally, we incorporate audio query Q_{audio} and learnable query Q_{learn} as a mixed query Q_{mixed} for the input of the transformer decoder.

The addition of learnable queries enhances our model's adaptability for various AVS tasks and datasets. It enables the model to learn dataset-level contextual information, and adjust the attention allocated to different sounding targets.

Transformer Encoder

The transformer encoder is responsible for building the mask feature. Specifically, we collect the backbone features of three resolutions (*i.e.*, 1/8, 1/16, and 1/32), and then flatten and concatenate them as the input queries for the transformer encoder. After that, the output features are reshaped to their original shapes, and the 1/8-scale features are taken out separately and 2× upsampled. Finally, we add the upsampled features to the 1/4-scale features from the visual backbone and obtain the mask feature $\mathcal{F}_{\text{mask}} \in \mathbb{R}^{T \times D \times h \times w}$, where $h = \frac{H}{4}, w = \frac{W}{4}$, and D is the embed dimension.

Audio-Visual Mixer

As illustrated in Figure 3, the segmentation mask is generated based on the mask feature, which plays a crucial role in the final results. However, since the audio semantics can vary widely, a static network may not be able to capture all of the relevant information. This limitation may hinder the model's ability to identify inconspicuous sounding objects.

To address this issue, we propose an audio-visual mixer as shown in Figure 4(b). The design of this module is based on channel attention, which allows the model to selectively amplify or suppress different visual channels depending on the audio feature, improving its ability to capture complex audio-visual relationships. Specifically, the mixer learns a set of weights ω through audio-visual cross-attention, and applies them to highlight the relevant channels. The whole process can be represented as follows:

$$\begin{aligned}\omega &= \text{softmax}\left(\frac{\mathcal{F}_{\text{audio}}{\mathcal{F}_{\text{mask}}}^T}{\sqrt{D/n_{\text{head}}}}\right)\mathcal{F}_{\text{mask}}, \\ \hat{\mathcal{F}}_{\text{mask}} &= \mathcal{F}_{\text{mask}} + \mathcal{F}_{\text{mask}} \odot \omega.\end{aligned} \tag{2}$$

Here, $\mathcal{F}_{\text{audio}}$ and $\mathcal{F}_{\text{mask}}$ represent the input audio feature and the initial mask feature, and $\hat{\mathcal{F}}_{\text{mask}}$ denotes the mixed mask feature. n_{head} means the number of attention heads, which is set to 8 by default following common practice.

Transformer Decoder

The transformer decoder is designed to build potential sparse queries, and optimally match the visual features with corresponding queries. We utilize the mixed query Q_{mixed} as the input query and the multi-scale visual features as key/-value. As the decoding process continues, the output queries Q_{output} continuously aggregate with visual features, ultimately combining the auditory and visual modalities and containing various target information.

To generate the segmentation masks, we multiply the mask feature $\hat{\mathcal{F}}_{\text{mask}} \in \mathbb{R}^{T \times D \times h \times w}$ obtained from the audio-visual mixer with the mixed queries Q_{output} from the query generator. Then, an MLP is used to integrate different channels. Finally, the model predicts the mask $\mathcal{M}$ through a fully connected layer:

$$\mathcal{M} = \text{FC}(\hat{\mathcal{F}}_{\text{mask}} + \text{MLP}(\hat{\mathcal{F}}_{\text{mask}} \cdot Q_{\text{output}})). \tag{3}$$

Here, $\text{MLP}(\cdot)$ represents the MLP layer, and $\text{FC}(\cdot)$ means the fully connected layer. The output $\mathcal{M} \in \mathbb{R}^{T \times N_{\text{class}} \times h \times w}$ is the predicted segmentation mask, with the dimension N_{class} denotes the number of semantic classes.

Loss Function

Auxiliary mixing loss. With the introduction of the audio-visual mixer, our model has maintained great capability in dealing with plentiful audio semantics, but its robustness is still insufficient when facing complex scenes. Thus, we design a mixing loss $\mathcal{L}_{\text{mix}}$ to supervise the mixer, enabling the model to more accurately locate target objects. Specifically, we integrate all channels of the mask feature $\hat{\mathcal{F}}_{\text{mask}}$ through a linear layer and predict a binary mask. At the same time, we extract all foreground labels in the ground truth as a new binary label and calculate the Dice loss (Milletari, Navab, and Ahmadi 2016) between them.

Total loss. The loss function comprises two parts: IoU loss and mixing loss. The IoU loss $\mathcal{L}_{\text{IoU}}$ is calculated by comparing the final segmentation mask with the ground truth. Here, we use Dice loss (Milletari, Navab, and Ahmadi 2016) for supervision. Considering that in AVS tasks, the proportion of segmented objects occupying the entire image is relatively small, the model can better focus on the foreground and reduce interference from the background by using Dice loss. Thus, the total loss of our method is:

$$\mathcal{L} = \mathcal{L}_{\text{IoU}} + \lambda \mathcal{L}_{\text{mix}}. \tag{4}$$

Here, λ is a coefficient that controls the effect of the auxiliary loss. We set $\lambda = 0.1$ as it performs best.

Discussion

AVSegFormer adopts a framework for segmentation that resembles Mask2Former (Cheng et al. 2022). However, it distinguishes itself by tailoring enhancements specifically for AVS tasks, accommodating the input of multi-modal information, which is a capability not inherent in Mask2Former. We introduce dual-tower backbone networks to ensure comprehensive extraction of both visual and auditory features. Besides, we devise a novel dense audio-visual mixer and a sparse audio-visual decoder that empower the proposed AVSegFormer to leverage auditory cues effectively, resulting in enhanced segmentation performance. These tailored designs hold substantial implications for addressing the emerging challenges of the AVS tasks.

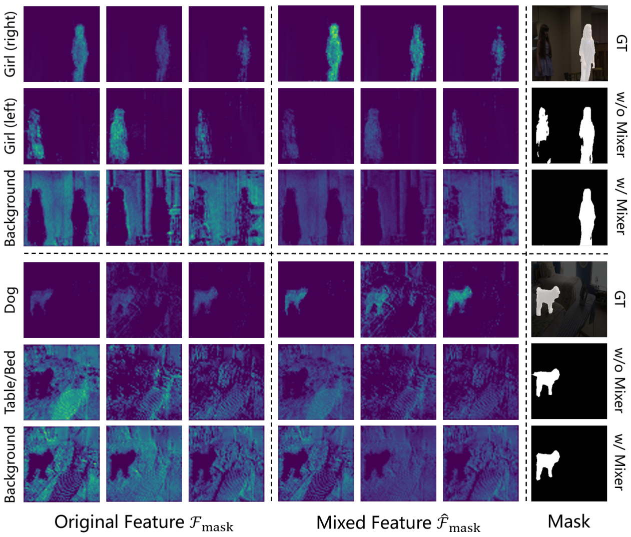


Figure 5: Comparison between the original features $\mathcal{F}_{\text{mask}}$ and the mixed features $\hat{\mathcal{F}}_{\text{mask}}$. We show two examples with 9 channels with the ground truth and predicted masks. As shown, the features of the ground truth (right girl and dog) are amplified while those of the non-sounding objects (left girl, table/bed, or background) are suppressed, which leads to different segmentation masks.

Experiments

Dataset

AVSBench-Object (Zhou et al. 2022) is an audio-visual dataset specifically designed for the audio-visual segmentation task, containing pixel-level annotations. The videos are downloaded from YouTube and cropped to 5 seconds, with one frame per second extracted for segmentation. The dataset includes two subsets: a single sound source subset for single sound source segmentation (S4), and a multi-source subset for multiple sound source segmentation (MS3). **S4 subset:** The S4 subset contains $4,932$ videos, with $3,452$ videos for training, 740 for validation, and 740 for testing. The target objects cover 23 categories, including humans, animals, vehicles, and musical instruments. **MS3 subset:** The MS3 subset includes 424 videos, with 286 training, 64 validation, and 64 testing videos, covering the same categories as the S4 subset.

AVSBench-Semantic (Zhou et al. 2023) is an extension of the AVSBench-Object, which offers additional semantic labels that are not available in the original AVSBench-Object dataset. It is designed for audio-visual semantic segmentation (AVSS). In addition, the videos in AVSBench-Semantic are longer, with a duration of 10 seconds, and 10 frames are extracted from each video for prediction. Overall, the AVSBench-Semantic dataset has increased in size by approximately three times compared to the original AVSBench-Object dataset, with 8,498 training, 1,304 validation, and 1,554 test videos.

Implementation Details

We train our AVSegFormer models for the three AVS subtasks using an NVIDIA V100 GPU. Consistent with previous works (Zhou et al. 2022, 2023), we employ AdamW (Loshchilov and Hutter 2017) as the optimizer, with a batch size of 2 and an initial learning rate of 2×10^{-5}. Since the MS3 subset is quite small, we train it for 60 epochs, while the S4 and AVSS subsets are trained for 30 epochs. The encoder and decoder in our AVSegFormer comprise 6 layers with an embedding dimension of 256. We set the coefficient of the proposed mixing loss $\mathcal{L}_{\text{mix}}$ to 0.1 for the best performance. More detailed training settings can be found in the supplementary materials.

Comparison with Prior Arts

We conducted a comprehensive comparison between our AVSegFormer and existing methods (Zhou et al. 2022, 2023; Mao et al. 2023) on the AVS benchmark. For fairness, we employ the ImageNet-1K (Deng et al. 2009) pre-trained ResNet-50 (He et al. 2016) or PVTv2 (Wang et al. 2022b) as the backbone to extract visual features, and the AudioSet (Gemmeke et al. 2017) pre-trained VGGish (Hershey et al. 2017) to extract audio features.

Method	Backbone	S4		MS3		AVSS		Reference
		F-score	mIoU	F-score	mIoU	F-score	mIoU	
LVS	ResNet-50	51.0	37.94	33.0	29.45	–	–	CVPR'2021
MSSL	ResNet-18	66.3	44.89	36.3	26.13	–	–	ECCV'2020
3DC	ResNet-34	75.9	57.10	50.3	36.92	21.6	17.27	BMVC'2020
SST	ResNet-101	80.1	66.29	57.2	42.57	–	–	CVPR'2021
AOT	Swin-B	–	–	–	–	31.0	25.40	NeurIPS'2021
iGAN	Swin-T	77.8	61.59	54.4	42.89	–	–	ArXiv'2022
LGVT	Swin-T	87.3	74.94	59.3	40.71	–	–	NeurIPS'2021
AVSBench-R50	ResNet-50	84.8	72.79	57.8	47.88	25.2	20.18	ECCV'2022
DiffusionAVS-R50	ResNet-50	**86.9**	75.80	62.1	49.77	–	–	ArXiv'2023
AVSegFormer-R50 (ours)	ResNet-50	85.9	**76.45**	62.8	49.53	29.3	24.93	AAAI'2024
AVSegFormer-R50* (ours)	ResNet-50	86.7	76.38	**65.6**	**53.81**	**31.5**	**26.58**	AAAI'2024
AVSBench-PVTv2	PVTv2	87.9	78.74	64.5	54.00	35.2	29.77	ECCV'2022
DiffusionAVS-PVTv2	PVTv2	90.2	81.38	70.9	58.18	–	–	ArXiv'2023
AVSegFormer-PVTv2 (ours)	PVTv2	89.9	82.06	69.3	58.36	42.0	36.66	AAAI'2024
AVSegFormer-PVTv2* (ours)	PVTv2	**90.5**	**83.06**	**73.0**	**61.33**	**42.8**	**37.31**	AAAI'2024

Table 1: Comparison with state-of-the-art methods on the AVS benchmark. All methods are evaluated on three AVS subtasks, including single sound source segmentation (S4), multiple sound source segmentation (MS3), and audio-visual semantic segmentation (AVSS). The evaluation metrics are F-score and mIoU. The higher the better. *We tried to enlarge the image resolution to 512×512.

N_{query}	S4		MS3	
	mIoU	F	mIoU	F
1	79.6	86.6	59.1	69.7
100	81.4	88.9	60.5	71.2
200	82.3	89.6	61.0	72.4
300	83.1	90.5	61.3	73.0

(a) Effect of the number of queries. We find that 300 queries work better than other settings.

learnable queries	S4		MS3	
	mIoU	F	mIoU	F
✓	83.1	90.5	61.3	73.0
×	82.7	89.9	58.5	70.9

(b) Effect of the learnable query. Using learnable queries along with audio queries improves the performance.

mixer	S4		MS3	
	mIoU	F	mIoU	F
–	81.4	88.0	59.4	70.9
CRA	82.7	89.7	59.8	72.2
CHA	83.1	90.5	61.3	73.0

(c) Effect of audio-visual mixer. It is shown that the channel-attention (CHA) mixer works better.

mix loss	S4		MS3	
	mIoU	F	mIoU	F
✓	83.1	90.5	61.3	73.0
×	81.3	89.1	59.6	71.3

(d) Effect of mixing loss. It shows that using mixing loss can indeed improve performance.

Table 2: AVSegFormer ablation experiments on the S4 and MS3 subsets. We report the performance of F-score (denoted as F) and mIoU. If not specified, the default settings are: the number of queries N_{query} is 300, the queries in the decoder are learnable, the audio-visual mixer is used, and the mixing loss is applied. Default settings are marked in gray.

Comparison with methods from related tasks. Firstly, we compare our AVSegFormer with state-of-the-art methods from three AVS-related tasks, including sound source localization (LVS (Chen et al. 2021) and MSSL (Qian et al. 2020)), video object segmentation (3DC (Mahadevan et al. 2020), SST (Duke et al. 2021) and AOT (Yang, Wei, and Yang 2021)), and salient object detection (iGAN (Mao et al. 2021) and LGVT (Zhang et al. 2021)). These results are collected from the AVS benchmark (Zhou et al. 2022), which are transferred from the original tasks to the AVS tasks.

As shown in Table 1, our AVSegFormer exceeds these methods by large margins. For instance, on the S4 subset, AVSegFormer-R50 achieves an impressive mIoU of 76.45, which is 1.51 points higher than the best LGVT. Although LGVT has a better Swin-T (Liu et al. 2021) backbone, our AVSegFormer with ResNet-50 backbone still performs better regarding mIoU. In addition, AVSegFormer-PVTv2 produces an outstanding mIoU of 82.06 and an F-score of 89.9 on this subset, which is 7.12 mIoU and 2.6 F-score higher than LGVT, respectively.

On the MS3 subset, AVSegFormer-R50 outperforms the best iGAN with 6.64 mIoU and 8.4 F-score, while AVSegFormer-PVTv2 further raised the bar with an exceptional improvement of 15.47 mIoU and 14.9 F-score. On the AVSS subset, our AVSegFormer-R50 yields 24.93 mIoU and 29.3 F-score, and AVSegFormer-PVTv2 obtains an impressive performance of 36.66 mIoU and 42.0 F-score, surpassing AOT by 11.26 mIoU and 11.0 F-score, respectively.

Comparison with AVSBench. Then, we compare our AVSegFormer with the AVSBench baseline, which is the current state-of-the-art method for audio-visual segmentation. As reported in Table 1, on the S4 subset, AVSegFormer-R50 achieves 3.66 mIoU and 1.1 F-score improvements over AVSBench-R50, while AVSegFormer-PVTv2 surpasses AVSBench-PVTv2 by 3.32 mIoU and 2.0 F-score. On the MS3 subset, AVSegFormer-PVTv2 surpasses AVSBench-PVTv2 with a margin of 1.65 mIoU and 5.0 F-score. On the AVSS subset, AVSegFormer-R50 and

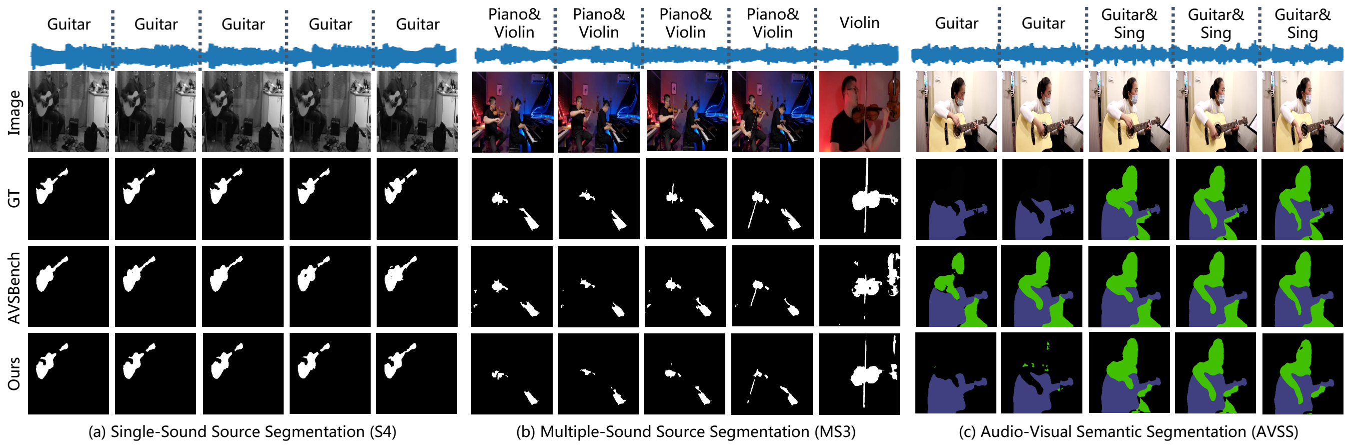


Figure 6: Qualitative results of AVSBench and AVSegFormer on three AVS sub-tasks. These results show that the proposed AVSegFormer can accurately segment the pixels of sounding objects and outline their shapes better than AVSBench.

AVSegFormer-PVTv2 achieve significant results with an mIoU improvement of 4.75 and 6.89, and a substantial F-score improvement of 4.1 and 6.8. These results demonstrate that AVSegFormer outperforms the AVSBench baseline on all sub-tasks, becoming a new state-of-the-art method for audio-visual segmentation.

Ablation Study

In this section, we conduct ablation experiments to verify the effectiveness of each key design in the proposed AVSegFormer. Specifically, we adopt PVTv2 (Wang et al. 2022b) as the backbone and conduct extensive experiments on the S4 and MS3 sub-tasks.

Number of queries. To analyze the impact of the number of queries on the model's performance, we conducted experiments with varying numbers of queries for the decoder input, specifically 1, 100, 200, and 300. Our results reveal a positive correlation between the number of queries and the model performance, with the optimal performance obtained when the number of queries was set to 300. Table 2a presents these findings.

Effect of learnable queries. We further investigated the impact of learnable queries in the decoder inputs. As shown in Table 2b, the improvement due to the learnable queries is relatively small in the single sound source task (S4), while it brings significant improvement in the multiple sound source task (MS3). This can be attributed to the complexity of sounding objects. We involve a more detailed discussion in the supplementary materials.

Effect of audio-visual mixer. We then studied the impact of the audio-visual mixer on our model. Two versions are designed for this module, as illustrated in Figure 4. The cross-attention mixer (CRA) utilizes visual features as queries and audio features as keys/values for cross-attention, and the channel-attention mixer (CHA) introduced the mechanism of channel attention with audio features as queries and visual features as keys/values. As presented in Table 2c, the design of CHA brought greater performance improvement compared to CRA.

In addition, we also visualize the mask feature before and after the audio-visual mixer, as shown in Figure 5. It is evident that for the sounding object (right girl and dog), the mixer effectively enhanced its features. Meanwhile, the non-sounding objects (left girl, table/bed, or background) experienced some degree of suppression. These findings align with our hypothesis and further substantiate the effectiveness of the audio-visual mixer.

Effect of auxiliary mixing loss. We finally conducted experiments to learn the impact of the auxiliary mixing loss. We train our model with and without the mixing loss, respectively, and report the testing results in Table 2d. It is demonstrated that the auxiliary mixing loss can help a lot in the final prediction.

Qualitative analysis. We also present the visualization results of AVSegFormer compared with those of AVSBench on three audio-visual segmentation tasks in Figure 6. The visualization results clearly demonstrate that our method performs better. It has a strong ability in target localization and semantic understanding and can effectively identify the correct sound source and accurately segment the target object in multiple sound source scenes. These results highlight the effectiveness and robustness of our method.

Conclusion

In this paper, we propose AVSegFormer, a novel audio-visual segmentation framework that leverages the power of transformer architecture to achieve leading performance. Specifically, our method comprises four key components: (1) a transformer encoder building the mask feature, (2) a query generator providing audio-conditioned queries, (3) a dense audio-visual mixer dynamically adjusting interested visual features, and (4) a sparse audio-visual decoder separating audio sources and matching optimal visual features. These components provide a more robust audio-visual cross-modal representation, improving the AVS performance in different scenarios. Extensive experimental results demonstrate the superior performance of AVSegFormer compared to existing state-of-the-art methods.

Acknowledgements

This work is supported by the Natural Science Foundation of China under Grant 62372223.

References

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, 2425–2433.

Arandjelovic, R.; and Zisserman, A. 2017. Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision*, 609–617.

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

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *Proceedings of the 16th European Conference of Computer Vision*, 213–229.

Chen, H.; Xie, W.; Afouras, T.; Nagrani, A.; Vedaldi, A.; and Zisserman, A. 2021. Localizing visual sounds the hard way. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16867–16876.

Chen, K.; Pang, J.; Wang, J.; Xiong, Y.; Li, X.; Sun, S.; Feng, W.; Liu, Z.; Shi, J.; Ouyang, W.; et al. 2019. Hybrid task cascade for instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4974–4983.

Chen, Z.; Duan, Y.; Wang, W.; He, J.; Lu, T.; Dai, J.; and Qiao, Y. 2022. Vision transformer adapter for dense predictions. In *International Conference on Learning Representations*.

Cheng, B.; Misra, I.; Schwing, A. G.; Kirillov, A.; and Girdhar, R. 2022. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1290–1299.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.

Deng, J.; Yang, Z.; Chen, T.; Zhou, W.; and Li, H. 2021. Transvg: End-to-end visual grounding with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1769–1779.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

Duke, B.; Ahmed, A.; Wolf, C.; Aarabi, P.; and Taylor, G. W. 2021. Sstvos: Sparse spatiotemporal transformers for video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5912–5921.

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 *2017 IEEE International Conference on Acoustics, Speech and Signal Processing*, 776–780.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, 2961–2969.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

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 *2017 IEEE International Conference on Acoustics, Speech and Signal Processing*, 131–135.

Jain, J.; Li, J.; Chiu, M.; Hassani, A.; Orlov, N.; and Shi, H. 2022. OneFormer: One Transformer to Rule Universal Image Segmentation. *arXiv preprint arXiv:2211.06220*.

Kamath, A.; Singh, M.; LeCun, Y.; Synnaeve, G.; Misra, I.; and Carion, N. 2021. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1780–1790.

Kirillov, A.; Girshick, R.; He, K.; and Dollár, P. 2019. Panoptic feature pyramid networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6399–6408.

Li, F.; Zhang, H.; Liu, S.; Zhang, L.; Ni, L. M.; Shum, H.-Y.; et al. 2022a. Mask dino: Towards a unified transformer-based framework for object detection and segmentation. *arXiv preprint arXiv:2206.02777*.

Li, Z.; Wang, W.; Xie, E.; Yu, Z.; Anandkumar, A.; Alvarez, J. M.; Luo, P.; and Lu, T. 2022b. Panoptic segformer: Delving deeper into panoptic segmentation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1280–1289.

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

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

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10012–10022.

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3431–3440.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

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. *arXiv preprint arXiv:2008.11516*.

Mao, Y.; Zhang, J.; Wan, Z.; Dai, Y.; Li, A.; Lv, Y.; Tian, X.; Fan, D.-P.; and Barnes, N. 2021. Transformer transforms salient object detection and camouflaged object detection. *arXiv preprint arXiv:2104.10127*.

Mao, Y.; Zhang, J.; Xiang, M.; Lv, Y.; Zhong, Y.; and Dai, Y. 2023. Contrastive Conditional Latent Diffusion for Audio-visual Segmentation. arXiv:2307.16579.

Milletari, F.; Navab, N.; and Ahmadi, S.-A. 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 40th International Conference on 3D Vision (3DV)*, 565–571.

Qian, R.; Hu, D.; Dinkel, H.; Wu, M.; Xu, N.; and Lin, W. 2020. Multiple sound sources localization from coarse to fine. In *Proceedings of the 16th European Conference on Computer Vision*, 292–308.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of the 18th International Conference of Medical Image Computing and Computer-Assisted Intervention*, 234–241.

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

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Wang, P.; Yang, A.; Men, R.; Lin, J.; Bai, S.; Li, Z.; Ma, J.; Zhou, C.; Zhou, J.; and Yang, H. 2022a. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, 23318–23340.

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

Wu, Q.; Wang, P.; Shen, C.; Dick, A.; and Van Den Hengel, A. 2016. Ask me anything: Free-form visual question answering based on knowledge from external sources. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4622–4630.

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

Xiong, Y.; Liao, R.; Zhao, H.; Hu, R.; Bai, M.; Yumer, E.; and Urtasun, R. 2019. Upsnet: A unified panoptic segmentation network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8818–8826.

Yang, Z.; Wei, Y.; and Yang, Y. 2021. Associating objects with transformers for video object segmentation. *Advances in Neural Information Processing Systems*, 34: 2491–2502.

Zhang, H.; Li, F.; Liu, S.; Zhang, L.; Su, H.; Zhu, J.; Ni, L. M.; and Shum, H.-Y. 2022. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*.

Zhang, J.; Xie, J.; Barnes, N.; and Li, P. 2021. Learning Generative Vision Transformer with Energy-Based Latent Space for Saliency Prediction. In *2021 Conference on Neural Information Processing Systems*.

Zhou, J.; Shen, X.; Wang, J.; Zhang, J.; Sun, W.; Zhang, J.; Birchfield, S.; Guo, D.; Kong, L.; Wang, M.; et al. 2023. Audio-Visual Segmentation with Semantics. *arXiv preprint arXiv:2301.13190*.

Zhou, J.; Wang, J.; Zhang, J.; Sun, W.; Zhang, J.; Birchfield, S.; Guo, D.; Kong, L.; Wang, M.; and Zhong, Y. 2022. Audio–Visual Segmentation. In *Proceedings of the European Conference on Computer Vision*, 386–403.

Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2020. Deformable DETR: Deformable Transformers for End-to-End Object Detection. In *International Conference on Learning Representations*.

Zhu, X.; Zhu, J.; Li, H.; Wu, X.; Li, H.; Wang, X.; and Dai, J. 2022. Uni-perceiver: Pre-training unified architecture for generic perception for zero-shot and few-shot tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16804–16815.