Segment beyond View: Handling Partially Missing Modality for Audio-Visual Semantic Segmentation

Renjie Wu, Hu Wang, Feras Dayoub, Hsiang-Ting Chen

The University of Adelaide

{renjie.wu, hu.wang, feras.dayoub, tim.chen}@adelaide.edu.au

Abstract

Augmented Reality (AR) devices, emerging as prominent mobile interaction platforms, face challenges in user safety, particularly concerning oncoming vehicles. While some solutions leverage onboard camera arrays, these cameras often have limited field-of-view (FoV) with front or downward perspectives. Addressing this, we propose a new out-of-view semantic segmentation task and Segment Beyond View (SBV), a novel audio-visual semantic segmentation method. SBV supplements the visual modality, which miss the information beyond FoV, with the auditory information using a teacher-student distillation model (Omni2Ego). The model consists of a vision teacher utilising panoramic information, an auditory teacher with 8-channel audio, and an audio-visual student that takes views with limited FoV and binaural audio as input and produce semantic segmentation for objects outside FoV. SBV outperforms existing models in comparative evaluations and shows a consistent performance across varying FoV ranges and in monaural audio settings.
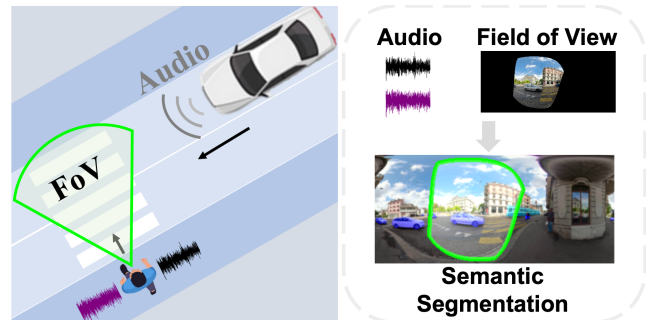
Figure 1: Pedestrian in the image can only see objects in the field of view (FoV) but hear the oncoming out-of-view vehicle and determine its general location and what kind of vehicle it is. The right image describes our novel task, with only FoV and binaural audio, the model can semantically segment the in- and out-of-view vehicles in the panorama.

Introduction

Over the span of 2009 to 2018 in Australia, there were 1,711 pedestrian fatalities and over 30,000 recorded hospitalizations stemming from injuries incurred in road accidents (Department of Infrastructure and Government 2018). These alarming statistics underline the growing concern over pedestrian safety in the era of mobile technology. The issue of pedestrian distraction may be further exacerbated by the rapid advancement of head-mounted displays (HMDs) and extended reality technologies. Despite all the potential benefits, HMDs compete for users' limited attentional resources, just like smart phones. Studies have shown that Augmented Reality (AR) application usage increases the risk of pedestrian hazards, such as colliding with obstacles or being unaware of approaching vehicles (Serino et al. 2016).

A few research works address the road safety issue by leveraging the cameras on mobile devices and HMDs. For examples, Wang et al. (2012) and Tong, Jia, and Bao (2021) used the onboard cameras to detect and predict vehicle trajectories and warn the user of potential collision. Kang, Lee, and Han (2019) presented a system for detecting the ground

obstacles along the path of pedestrians. Nonetheless, the focus on compactness and user comfort in mobile devices limit the placement of the camera system, resulting in a restricted field of view (FoV) that only marginally exceeding human vision. As a consequence, rendering both the user and the AI model blind to potential road hazards, which frequently originate from areas outside the current FoV. Researchers also explored the use of audio signal to infer the out-of-view objects (Manori et al. 2018; Mizumachi et al. 2014; Rovetta, Mnasri, and Masulli 2020), yet these approaches often lack precision in locating the position of upcoming vehicles.

Recognising this challenge, we introduce a new semantic segmentation task for objects beyond the field of view, with a benchmark that focuses on identifying oncoming vehicles for HMD users safety. This introduces a novel *partially missing modality problem*, where the model only has access to partial information within a specific modality, such as a constrained FoV or monaural audio signals to the surrounding environment. This problem formulation diverges from both the conventional *multi-modality problem*, which aims to enhance downstream tasks, such as tracking or segmentation accuracy, by leveraging multiple modalities (Valverde, Hurtado, and Valada 2021; Chakravarthula et al. 2023; Zürn and Burgard 2022; Kim and Kim 2023), and the *cross-*

modality problem, which focuses on the knowledge transfer from one modality to another in cases of absent information (Gan et al. 2019; Dai et al. 2022). In contrast, our focus is specifically on situations where the modality is only partially missing, which provides the opportunity of utilizing the available signal in conjunction with data from other modalities. It should be noted that the partially missing modality problem can be seen as a specific, yet significant, case within the cross-modality problem, where one modality is partially missing (e.g., limited FoV vs panorama).

To tackle this task, we propose Segment Beyond View (SBV), an audio-visual semantic segmentation method that supplements the visual modality, which partially miss the information beyond FoV, with the auditory information. SBV is driven by a teacher-student distillation model, which we termed *Omni2Ego*, comprising a vision teacher utilising panoramic information, an auditory teacher with 8-channel audio, and an audio-visual student that takes views with limited FoV and binaural audio as input and produce semantic segmentation for objects outside the FoV. Fig. 1 shows the illustration of our task. Adapting the Omni Auditory Perception Dataset (Dai et al. 2022; Vasudevan, Dai, and Van Gool 2020) to the proposed task, the results suggest that our method outperforms state-of-the-art audio-visual semantic segmentation methods (Zhou et al. 2022, 2023) and maintain consistent performance across different FoV ranges and in monaural audio environments.

Our work makes following contributions: **(1)** Presenting a simple yet effective framework, termed Segment Beyond View (SBV), that leverages the partially-available information in one modality and complements it with information from another modality to perform the out-of-view semantic segmentation; **(2)** Introducing a novel out-of-view semantic segmentation task and its associated benchmark based on public dataset; **(3)** Demonstrating the superior performance of SBV through comparison with state-of-the-art models and presenting ablation studies examining various degrees of partially missing modality and model architectures. Additionally, our task has potential implications to robot navigation, autonomous vehicles and road safety in general.

Related Work

Pedestrian Safety

Numerous studies have found that engaging in activities such as texting, talking, playing AR games on a phone can lead to unsafe street crossing behaviours, such as delayed initiation of street crossing, stepping onto the street before vehicles come to a complete stop, and slower walking pace while crossing the street (Campisi et al. 2022; Serino et al. 2016; Cortes et al. 2023). These research also suggests that pedestrians may underestimate the risks associated with their distracted behaviour, further exacerbating the problem.

Many works have been dedicated to enhancing pedestrian safety through the detection of road hazards and guiding users' attention towards them. For example, previous studies used smartphone cameras (Wang et al. 2012; Holländer, Krüger, and Butz 2020) to detect potential hazards in the vicinity. Novel visualization methods have also been ex-

plored to notify users of out-of-view objects, such as Halo and Wedge (Gruenefeld et al. 2018), EyeSee360 (Gruenefeld et al. 2017), and EdgeRadar (Gustafson and Irani 2007). Jung et al. (2018) integrated these visualisation methods with a vehicle position estimation model based on an additional wearable camera. Similarly, some AR approaches (Tong, Jia, and Bao 2021; Wu and Chen 2023) are developed to inform pedestrians of the collision direction. All these works rely on the assumption that cameras could capture approaching vehicles, while our work alleviate this limitation and demonstrate the potential of identifying oncoming vehicles outside the FoV.

Multimodal Learning with Missing Modality

Multimodal learning with missing modalities has gained much attention recently. Some methods aim to make predictions even when some modalities are unavailable during training or testing. Some approaches, such as those by Recasens et al. (2023) and Ma et al. (2022) apply masks or optimize multi-task strategies to handle missing modalities. Other methods handle missing modalities by predicting weights (Miech, Laptev, and Sivic 2018) or using combinatorial loss (Shvetsova et al. 2022). A recent work (Li, Liu, and Tang 2022) proposes an audio-visual tracker that can localize speaker targets in the absence of visual modality. However, those methods require modality-complete training data. The SMIL (Ma et al. 2021) and ShaSpec (Wang et al. 2023) are developed specifically for handling multimodal learning with missing modalities both during training and testing. But above methods all assume that one or more modalities will be missing entirely instead of partially missing of certain modalities. However, our partially missing modality task assumes all kinds of modalities exist, but each of them is partially missing.

Audio-Visual Segmentation

Many audio-visual tasks have been proposed in recent years, such as visual sound source localization (Hu et al. 2020; Kamath et al. 2021) and audio-visual event localization (Tian et al. 2018; Wu et al. 2019). Existing audio-visual segmentation methods (Chen et al. 2021; Liu et al. 2022) have made significant progress but cannot distinguish object categories. AVSBench and TPAVI (Zhou et al. 2022) for audio-visual semantic segmentation is recently proposed to address this issue. But all above segmentation methods tend to locate visibly sound-making objects, the auditory signal also includes out-of-view sound-making objects. Those invisible out-of-view sound-making objects are also important, especially for AR user safety. The invisible part in our task setting are the missing part of visual modality. In this paper, along with a partially missing modality task and its settings, we propose a Segment Beyond View (SBV) model to tackle the partially missing modalities issues existing in multi-modal learning.

Method

Problem Definition

We are interested in audio-visual semantic segmentation with partially missing modality. Given an audiovisual
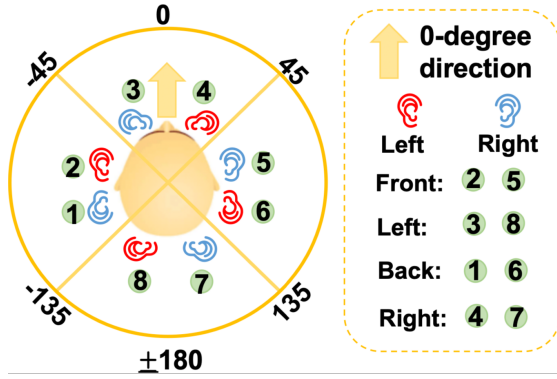
Figure 2: Description of binaural audio selected by the left and right rotation of the head. The number of left turns negative, and the number of right turns positive.

dataset contains omnidirectional auditory or visual information: panoramas and binaural audios in four directions (front, back, left, and right). Panoramas and all audios are accessible during training, but only partial view of the panoramas and the binaural audios in one direction are available in testing. Also, the training data contains no manual annotations. Our scenario is under pedestrian road safety, we define the partially visible view as the first-person view of the pedestrians. Formally, given a dataset $\mathcal{D} = \{\mathcal{D}_a, \mathcal{D}_v\}$, where $\mathcal{D}_a$ means the auditory modality part and $\mathcal{D}_v$ means the visual modality part. $\mathcal{D}_a$ contains binaural audio from 4 directions and $\mathcal{D}_v$ contains panoramas. For each panorama in the dataset, a randomly generated head rotation is assigned and the corresponding binaural audio is selected. We follow the previous works (Gao and Grauman 2019; Dai et al. 2022; Gao et al. 2020) to transfer the selected binaural audio to spectrograms, and denote $\mathbf{x}^{\text{isp}}$ and $\mathbf{y}^{\text{dsp}}$ as the spectrograms of input audio signals (isp) and spectrograms of their difference (dsp) with other directions. We denote $\mathbf{x}^{\text{osp}}$ as the spectrograms of audio in other directions. We denote the First-Person View (FPV) generated based on FoV and head rotations as $\mathbf{x}^{\text{fpv}}$ for the corresponding panoramic image $\mathbf{x}^v$. Our task is to use such datasets to train a model that can semantically segment the vehicles in the surrounding environment when the FPV and binaural audio are available.

Partially Missing Settings The "partially missing" settings have visual and auditory parts. The visible part of the panorama is the first-person view generated by the head rotation. For visual modality, the "partially missing" is the out-of-view scene. The binaural audio is selected according to the head rotation. For auditory modality, the "partially missing" is the binaural audio from other directions. Head rotation consists of three parameters: horizontal, vertical, and in-plane rotation. Formally, $\text{u} \in (-180°, 180°)$, $\text{v} \in (-90°, 90°)$, $\text{rot} \in (-180°, 180°)$, where, "u", "v" and "rot" represent the horizontal, vertical and in-plane rotation viewing angles, respectively. We visualize the binaural

sound selection process in the Fig. 2. Formally,

$$\mathbf{F}(\mathbf{u}) = \begin{cases} \{2, 5\}, & \mathbf{u} \in (-45°, \quad 45°) \\ \{4, 7\}, & \mathbf{u} \in (\quad 45°, \quad 135°) \\ \{3, 8\}, & \mathbf{u} \in (-45°, -135°) \\ \{1, 6\}, & \text{otherwise} \end{cases} \quad (1)$$

where, $\{1, .., 8\}$ means the id number of microphones and $\mathbf{F}(\cdot)$ denotes the mapping function for the id numbers in Fig. 2. Regarding the first-person view, we opt for the binocular overlap area size equivalent to human eyes, measuring $135°$ vertically and $120°$ horizontally (Wandell 1995), as this is crucial for comprehensive environmental perception.

Sound-Making Objects Extraction We use the following steps to generate foreground masks ($\mathbf{M}^{\text{fg}}$) for sound-making vehicles, etc., which are in line with the previous work (Dai et al. 2022). We use GSoC algorithm (Samsonov 2017) in OpenCV (Bradski 2000) to extract video backgrounds instead of the simple one mentioned in the previous work. Given a panoramic image ($\mathbf{x}^v$) and a background ($\mathbf{x}^{\text{bg}}$), first use the semantic segmentation model pre-trained on the Cityscapes dataset (Cordts et al. 2016) to get their semantic segmentation results: $\mathbf{y}^{\text{seg}}$ and $\mathbf{y}^{\text{bg}}$. $\mathbf{M}^{\text{fg}}$ is generated using the following formula:

$$\mathbf{M}^{\text{fg}}_{(h,w)} = \begin{cases} 1, & \text{if } \mathbf{y}^{\text{seg}}_{(h,w)} \in \{\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3\} \text{ and } \mathbf{y}^{\text{seg}}_{(h,w)} \neq \mathbf{y}^{\text{bg}}_{(h,w)} \\ 0, & \text{otherwise} \end{cases}$$

$$(2)$$

where (h, w) denotes the coordinates of the pixels, 1 is to keep the pixel that is belong to the sound-making vehicles and 0 means otherwise. $\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3$ mean car, tram and motorcycle correspondingly. We achieve similar results (mIoU: 65.39%) on *AuditoryTestManual* dataset (see in Section "Dataset") with the previous work (mIoU: 65.35%) using the same DeepLabv3+ (Chen et al. 2018) framework. Therefore, we consider our method can successfully generate the foreground masks, as some results shown by Fig. 5.

Model Architecture

To the best of our knowledge, we are the first to solve partially missing modality for audio-visual semantic segmentation and we also propose a new strong baseline. We adopt teacher-student distillation framework to train our encoder-decoder based model: Segment Beyond View (SBV). Fig. 3 shows the overall training architecture. For visual and auditory data, we employ two encoders respectively. Specifically, $\mathbf{f}^{\text{enc}}_{\text{vs}} = f_{\theta^{\text{enc}}_{\text{vs}}}(\mathbf{x}^{\text{fpv}})$, where, $f_{\theta^{\text{enc}}_{\text{vs}}}$ means the visual student encoder parameterized by $\theta^{\text{enc}}_{\text{vs}}$, $\mathbf{f}^{\text{enc}}_{\text{vs}}$ represents the feature ouputted by the visual encoder; similarly, $\mathbf{f}^{\text{enc}}_{\text{as}} = f_{\theta^{\text{enc}}_{\text{as}}}(\mathbf{x}^{\text{isp}})$, where $f_{\theta^{\text{enc}}_{\text{as}}}$ is the auditory encoder, "as" means the auditory student encoder and $\mathbf{f}^{\text{enc}}_{\text{as}}$ is the output feature.

Audio-Visual Feature Fusion Module (AVFFM) For our partially missing modality task, only first-person view is visible, and sound-making objects in the invisible out-of-view area need to be located by audio. We adopt an attention module for concatenated cross-modal features to better capture semantic-rich information with not only the multi-modal attentions, but also the uni-modal attentions. For details, we sent the visual feature map $\mathbf{f}^{\text{enc}}_{\text{vs}}$ and the audio feature map $\mathbf{f}^{\text{enc}}_{\text{as}}$ to the AVFFM. Since the $\mathbf{f}^{\text{enc}}_{\text{as}}$ has different
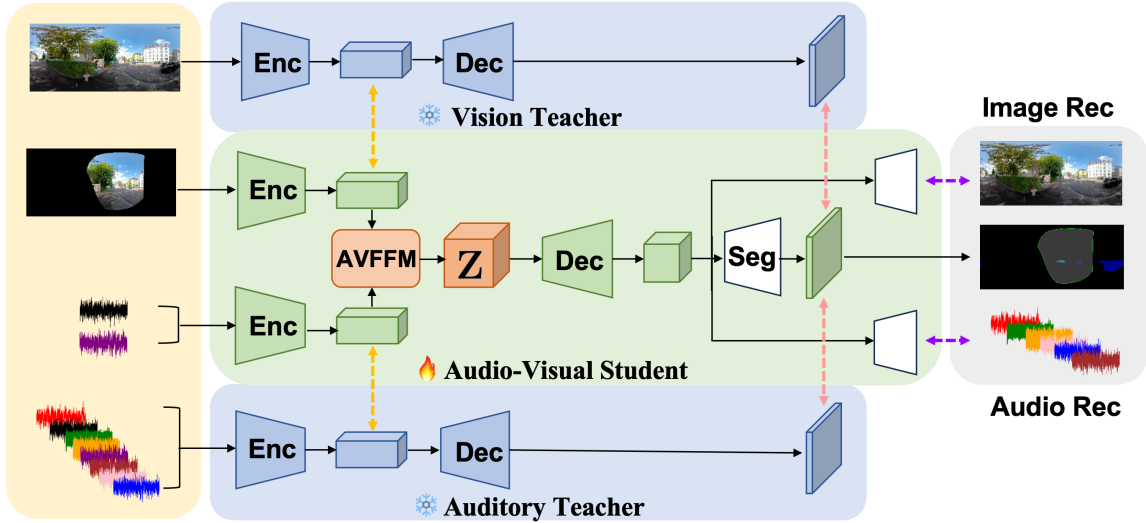
Figure 3: Segment Beyond View training architecture consists of a vision teacher, an auditory teacher, and an audio-visual student. The student takes first-person view and binaural audio as inputs. The input of the vision teacher is panoramas and input of the auditory teacher is the 8-channel audio. Enc: Encoder; Dec: Decoder; Seg: Segmentation Head; Rec: Reconstruction.

feature map size with the $\mathbf{f}_{\mathrm{vs}}^{\mathrm{enc}}$, we perform a following alignment process: $\tilde{\mathbf{f}}_{\mathrm{as}}^{\mathrm{enc}} = f_{\theta^{\mathrm{rsz}}}(\mathbf{f}_{\mathrm{as}}^{\mathrm{enc}})$, where, $f_{\theta^{\mathrm{rsz}}}$ means the resize operation that used to align the size and $\tilde{\mathbf{f}}_{\mathrm{as}}^{\mathrm{enc}}$ is the updated feature of the auditory encoder. Then we concatenate them and send it to do the dot-product measured attention (Zhou et al. 2022). Formally, for each $\mathbf{f}_{\mathrm{vs}}^{\mathrm{enc}} \in \mathbb{R}^{H \times W \times C}$ and $\tilde{\mathbf{f}}_{\mathrm{as}}^{\mathrm{enc}} \in \mathbb{R}^{H \times W \times C'}$, we get the audio-visual feature map $\mathbf{F}_{\mathrm{av}} \in \mathbb{R}^{H \times W \times (C+C')}$:

$$
\begin{aligned}
\mathbf{Q} &= \omega(\mathbf{F}_{\mathrm{av}}), \ \mathbf{K} = \phi(\mathbf{F}_{\mathrm{av}}), \ \mathbf{V} = g(\mathbf{F}_{\mathrm{av}}), \\
\mathbf{Z} &= \mathbf{F}_{\mathrm{av}} + \mu(\frac{\mathbf{Q}\mathbf{K}^T}{\mathbf{N}}\mathbf{V}),
\end{aligned}
\tag{3}
$$

where, $\mathbf{Z} \in \mathbb{R}^{H \times W \times (C+C')}$ denotes the updated audio-visual feature map, $\mathbf{Q}$, $\mathbf{K}$ and $\mathbf{V}$ represent the `query`, `key` and `value` in the attention mechanism, ϕ, ω, g and μ are all 1×1 convolution layers, $\mathbf{N} = H \times W$ and serves as a normalization factor.

Our decoder contains 4 stages. Formally, we denote size of each stage's output feature map as $\mathbf{F}_g^{\mathrm{dec}} \in \mathbb{R}^{H_g \times W_g \times C_g}$, where $(H_g, W_g) = \frac{(H^{\mathrm{img}}, W^{\mathrm{img}})}{2^{g+1}}, g = \{0, 1, 2, 3\}$ and $C_g = 512$. The segmentation head restores the feature map size and classifies each pixel. "$(H^{\mathrm{img}}, W^{\mathrm{img}})$" means the original height and width of input images. We denote the process of the decoder as: $\mathbf{F}_s^{\mathrm{dec}} = f_{\theta_s^{\mathrm{dec}}}(\mathbf{Z})$, where $\mathbf{F}_s^{\mathrm{dec}}$ is the output of the decoder (equivalent to $\mathbf{F}_g^{\mathrm{dec}}$ at $g = 0$) and $f_{\theta_s^{\mathrm{dec}}}$ represents our decoder. For details, $\mathbf{f}_s^{\mathrm{seg}} = f_{\theta_s^{\mathrm{seg}}}(\mathbf{F}_s^{\mathrm{dec}})$, where, $f_{\theta_s^{\mathrm{seg}}}$ denotes the segmentation head of the student model and $\mathbf{f}_s^{\mathrm{seg}} \in \mathbb{R}^{H^{\mathrm{img}} \times W^{\mathrm{img}} \times K}$ is the output logits of the segmentation head. K is the number of classes. After that, we use the softmax operation to get the segmentation results.

Audio and Image Reconstruction Head: To make the model have the ability to reconstruct the features of the sound-making objects in the out-of-view area and the richer position information of those objects in the audio information from other directions at the feature level, we introduce the image and audio reconstruction tasks as auxiliary tasks. Image reconstruction head restores the size to the input size. For details, $\mathbf{f}_s^{\mathrm{v}} = f_{\theta_s^{\mathrm{v}}}(\mathbf{F}_s^{\mathrm{dec}})$, where, $g = 3$, $f_{\theta_s^{\mathrm{v}}}$ means the image reconstruction head and $\mathbf{f}_s^{\mathrm{v}} \in \mathbb{R}^{H^{\mathrm{img}} \times W^{\mathrm{img}} \times 3}$ is the reconstructed image. Following the previous work (Gao and Grauman 2019) for the audio reconstruction head, the model tries to predict the differences between each channel audio. We use left ear audio to predict left ear audio in other directions, vice versa. We use the following steps to get the spectrograms of the differences between each channel audio: $\tilde{\mathbf{y}}^{\mathrm{dsp}} = f_{\theta_s^{\mathrm{a}}}(\mathbf{F}_s^{\mathrm{dec}})$, where, $g = 3$, $\tilde{\mathbf{y}}^{\mathrm{dsp}}$ is the predicted spectrogram of the differences and $f_{\theta_s^{\mathrm{a}}}$ is the audio reconstruction head and some post-process methods.

Omni2Ego Distillation

Knowledge Distillation (KD) attempts to preserve useful knowledge from the teacher into the student as the teacher can acquire more information than the student during training. While in testing, student cannot access full modalities information, KD can thus help to improve the performance. We propose "*Omni2Ego*" distillation method which is extremely simple but effective and can distill **Omni**-directional information in**to** the **Ego**-centric perspective, in both the visual and auditory aspects. Specifically, we distill panoramic visual information into the first-person view for completing missing visual part at the feature level. We also distill 8-channel audio information into 2-channel binaural audio for completing missing auditory information from other directions. We choose encoder-decoder architecture based SegFormer (Xie et al. 2021) and 8-channel SoundNet (Dai et al. 2022) as our visual and auditory teachers. Our method is divided into feature alignment and logits distillation parts.

Feature Alignment Distillation We denote the outputs of

the encoders of the visual and auditory teachers as: $\mathbf{r}_{\text{vt}}^{\text{enc}}$ and $\mathbf{f}_{\text{at}}^{\text{enc}}$, respectively. "at" and "vt" present the auditory teacher and visual teacher. Formally, $\mathbf{r}_{\text{vt}}^{\text{enc}} = f_{\theta_{\text{vt}}^{\text{enc}}}(\mathbf{x}^{\text{v}})$ and $\mathbf{f}_{\text{at}}^{\text{enc}} = f_{\theta_{\text{at}}^{\text{enc}}}(\mathbf{x}^{\text{isp}}, \mathbf{x}^{\text{osp}})$, where $f_{\theta_{\text{vt}}^{\text{enc}}}$ and $f_{\theta_{\text{at}}^{\text{enc}}}$ represent the encoders of visual and auditory teachers. "$(\mathbf{x}^{\text{isp}}, \mathbf{x}^{\text{osp}})$" is the 8-channel audio spectrograms. We use a linear layer and an interpolation operation to align the size of $\mathbf{r}_{\text{vt}}^{\text{enc}}$ with the size of $\mathbf{f}_{\text{vs}}^{\text{enc}}$ during training. Formally, $\mathbf{f}_{\text{vt}}^{\text{enc}} = f_{\theta^{\text{proj}}}(\mathbf{r}_{\text{vt}}^{\text{enc}})$, where $f_{\theta^{\text{proj}}}$ is the projection layer that can align the feature map size and $\mathbf{f}_{\text{vt}}^{\text{enc}}$ represents the aligned feature map of the visual teacher encoder. We abandon the $f_{\theta^{\text{proj}}}$ during testing.

Logits Distillation For the logits from visual and auditory teachers, we denote them as: $\mathbf{r}_{\text{vt}}^{\text{seg}}$ and $\mathbf{f}_{\text{at}}^{\text{seg}}$, respectively. To be specific, $\mathbf{r}_{\text{vt}}^{\text{seg}} = f_{\theta_{\text{vt}}^{\text{seg}}}(\mathbf{r}_{\text{vt}}^{\text{enc}})$ and $\mathbf{f}_{\text{at}}^{\text{seg}} = f_{\theta_{\text{at}}^{\text{seg}}}(\mathbf{f}_{\text{at}}^{\text{enc}})$, where $f_{\theta_{\text{vt}}^{\text{seg}}}$ and $f_{\theta_{\text{at}}^{\text{seg}}}$ denote the segmentation head of the visual and auditory teachers. To target logits from the visual teacher, we apply foreground masks $\mathbf{M}^{\text{fg}}$ from the Eqn. 2. Specifically, $\mathbf{f}_{\text{vt}}^{\text{seg}} = \mathbf{M}^{\text{fg}}\mathbf{r}_{\text{vt}}^{\text{seg}}$, where $\mathbf{f}_{\text{vt}}^{\text{seg}}$ means the target logits.

Training Objectives

The objective function for training is divided into three parts: Feature Alignment Loss, Logits Distillation Loss and Modality Reconstruction Loss, as shown in Fig. 3.

Feature Alignment Loss (FAL) FAL is based on the L2 loss and is divided into visual and auditory parts:

$$\ell_{\text{fal}}(\mathcal{D}, \Theta^{\text{enc}}) = \sum_{i=1}^{|\mathcal{D}|} \|\mathbf{f}_{\text{at}}^{\text{enc}} - \mathbf{f}_{\text{as}}^{\text{enc}}\|_2 + \sum_{i=1}^{|\mathcal{D}|} \|\mathbf{f}_{\text{vt}}^{\text{enc}} - \mathbf{f}_{\text{vs}}^{\text{enc}}\|_2, \quad (4)$$

where, $\ell_{\text{fal}}(\cdot)$ means the feature alignment operator, $\|\cdot\|_2$ means the two norm operator, $\Theta^{\text{enc}} = \{\theta_{\text{vt}}^{\text{enc}}, \theta_{\text{vs}}^{\text{enc}}, \theta_{\text{at}}^{\text{enc}}, \theta_{\text{as}}^{\text{enc}}\}$.

Logits Distillation Loss (LDL) LDL is divided into visual and auditory parts. Since we only focus on three categories of moving sound-making objects, in order to make the student model pay more attention to the important features from the teacher model, we use L1 loss for logits distillation:

$$\ell_{\text{ldl}}(\mathcal{D}, \Theta^{\text{seg}}) = \sum_{i=1}^{|\mathcal{D}|} \|\mathbf{f}_{\text{at}}^{\text{seg}}, \mathbf{f}_{\text{s}}^{\text{seg}}\|_1 + \sum_{i=1}^{|\mathcal{D}|} \|\mathbf{f}_{\text{vt}}^{\text{seg}}, \mathbf{f}_{\text{s}}^{\text{seg}}\|_1, \quad (5)$$

where, $\ell_{\text{ldl}}(\cdot)$ denotes the overall logits distillation operator, $\|\cdot\|_1$ denotes the L1 loss operator, $\Theta^{\text{seg}} = \{\theta_{\text{at}}^{\text{seg}}, \theta_{\text{vt}}^{\text{seg}}, \theta_{\text{s}}^{\text{seg}}\}$.

Modality Reconstruction Loss (MRL) MRL has two parts, one is to reconstruct the panoramic image, another is to reconstruct the binaural audio from other directions:

$$\ell_{\text{rec}}(\mathcal{D}, \Theta^{\text{rec}}) = \sum_{i=1}^{|\mathcal{D}|} \|\mathbf{y}^{\text{dsp}} - \tilde{\mathbf{y}}^{\text{dsp}}\|_2 + \sum_{i=1}^{|\mathcal{D}|} \|\mathbf{x}^{\text{v}} - \mathbf{f}_{\text{s}}^{\text{v}}\|_2, \quad (6)$$

where, $\ell_{\text{rec}}(\cdot)$ denotes the reconstruction operator, $\Theta^{\text{rec}} = \{\theta_{\text{s}}^{\text{a}}, \theta_{\text{s}}^{\text{v}}\}$. We use L2 loss here.

Overall Loss (OL) Below give the overall loss $\ell_{\text{task}}(\cdot)$:

$$\ell_{\text{task}}(\mathcal{D}, \Theta) = \lambda\ell_{\text{fal}}(\mathcal{D}, \Theta^{\text{enc}}) + \beta\ell_{\text{ldl}}(\mathcal{D}, \Theta^{\text{seg}}) + \gamma\ell_{\text{rec}}(\mathcal{D}, \Theta^{\text{rec}}), \quad (7)$$

where, $\Theta = \{\Theta^{\text{seg}}, \Theta^{\text{enc}}, \Theta^{\text{rec}}\}$, $\lambda = \{\lambda_a, \lambda_v\}$, $\beta = \{\beta_a, \beta_v\}$ and $\gamma = \{\gamma_a, \gamma_v\}$. λ_a and λ_v are trade-off factors used to mediate between auditory and visual features. β_a and β_v are trade-off factors for auditory and visual logits. γ_a and γ_v are coefficients for audio and image reconstruction.

Experiments

Dataset

Existing omnidirectional audio-visual semantic segmentation datasets for road safety are limited. **Omni Auditory Perception Dataset** (Dai et al. 2022; Vasudevan, Dai, and Van Gool 2020) is a dataset that contains $64, 250$ 2-second video clips with 8 channel audio of city traffic in Zurich that are recorded by a $360°$ GoPro Fusion cameras and 4 pairs of 3Dio binaural microphones in four directions (front, back, left and right). In addition to the normal training dataset $(51, 400)$ and validation dataset $(6, 208)$, it contains two test datasets: *AuditoryTestPseudo* dataset $(6, 492)$ and *AuditoryTestManual* dataset. The annotations for *AuditoryTestPseudo* dataset are generated by the model pre-trained on the Cityscapes dataset (Cordts et al. 2016). Most objects that need to be segmented are in the equatorial region of the panorama without obvious distortion (Dai et al. 2022). Therefore, those pre-trained models can achieve satisfactory performances. *AuditoryTestManual* dataset is a manually labeled dataset contains a total of 80 images but covers a variety of scenarios including rainy, foggy, night and daylight. This dataset contains three categories: car, tram, and motorcycle. Each sample has a panorama that is the middle frame of the video clip and eight 2-second audio clips (this setting follows previous works (Gan et al. 2019; Dai et al. 2022)).

Implementation Details

We train models by using NVIDIA A100 GPUs. We use Adam (Kingma and Ba 2014) and set learning rate as 1×10^{-5} for the optimizer. We use one cycle policy (Smith and Topin 2019) as our learning rate decay strategy. All images are resized to 480×480. The spectrogram size is set as 257×601. All student models are trained for 50 epochs to ensure that the loss converges. For the Eqn. 7, we set $\beta_a = 0.1$ and $\beta_v = 0.4$ for logits distillation; about the feature distillation part, we set all $\lambda = 0.05$ and all $\gamma = 0.02$.

We choose SegFormer (Xie et al. 2021) pretrained on the Cityscapes dataset and 8-channel SoundNet (Dai et al. 2022) as teacher models. For student visual encoder, we followed previous work (Zhou et al. 2022) and chose ResNet50 (He et al. 2016) with an Atrous Spatial Pyramid Pooling (ASPP) module (Chen et al. 2018); about auditory encoder which is same with the SoundNet's encoder. The segmentation head consists of three convolution layers and one interpolation operation. The components of the image reconstruction head are three convolution layers and one upsampling layer. The audio reconstruction head has 5 convolution layers.

Overall Performance

Following previous works (Vasudevan, Dai, and Van Gool 2020; Dai et al. 2022; Zhou et al. 2022), we present the F_β-score ($\beta = 0.3$) and mean Intersection-over-Union (mIoU) of the following baseline methods and our models in the Tab. 1, since our task is still an audio-visual semantic segmentation task. Our task divides the panorama into first-person view and out-of-view area, we apply the above two metrics to both area, which is simple to realize and just apply

Methods	Input	AuditoryTestPseudo						AuditoryTestManual					
		FPV		OOV		All		FPV		OOV		All	
		FS	mIoU	FS	mIoU	FS	mIoU	FS	mIoU	FS	mIoU	FS	mIoU
SF	Pano	-	-	-	-	-	-	-	-	-	-	.820	67.3
SN	8-C	-	-	-	-	.588	42.5	-	-	-	-	.547	40.4
SF	FPV	.755	69.1	.207	12.7	.440	30.4	.688	62.7	.186	11.4	.396	27.2
SBV-V	FPV	.758	69.5	.264	16.2	.471	32.5	.700	63.9	.226	14.0	.427	29.4
SN	2-C	-	-	-	-	.501	36.2	-	-	-	-	.458	32.3
SBV-A	2-C	-	-	-	-	.545	38.4	-	-	-	-	.491	35.3
TPAVI	FPV & 2-C	.795	71.6	.505	39.0	.647	47.5	.754	69.1	.410	31.9	.586	42.5
SBV	**FPV & 2-C**	**.817**	**73.8**	**.590**	**46.7**	**.705**	**53.0**	**.777**	**70.3**	**.551**	**43.3**	**.679**	**50.1**

Table 1: The general table contains the comparison with other baselines. Pano: Panorama: 2-C/8-C: 2-Channel/8-Channel; FPV: first-person view; OOV: out-of-view; SF: SegFormer (Xie et al. 2021); SN: SoundNet (Dai et al. 2022); TPAVI (Zhou et al. 2022). Here we report mIoU (%) and F-score (FS). It should be noted that we fill some cells with "-", because for the models that take panorama and audio as inputs do not have first-person view or out-of-view and we only report the overall performance.

first-person view and out-of-view masks on segmentation results and ground truth and then to do the evaluation.

Vision Models: We also choose the SegFormer (Xie et al. 2021) with only first-person view input to verify that it is impossible to achieve satisfactory performance with visual inputs alone. For "*SBV-V*", this is an variant of our model, we disable the auditory encoder, AVFFM and image reconstruction and only apply visual feature alignment and logits distillation. *SBV-V* with only first-person view input can also achieve higher performance than SegFormer with only first-person view input by using our visual distillation method from panorama to first-person view, and resulted in average increases of 3.5 / 2.1% in mIoU over out-of-view / overall.

Auditory Models: We use the 2-channel SoundNet (Dai et al. 2022) as the auditory input only method and it is used to show panoramic semantic segmentation using only binaural audio is challenging. For "*SBV-A*", it is another variant of our SBV, we disable the visual encoder, AVFFM and audio reconstruction and only apply auditory feature and logits distillation. We can see *SBV-A* with only binaural audio input also outperforms 2-channel SoundNet by using our auditory distillation method by about 2.2 / 3.2% mIoU on two *AuditoryTestPseudo* and *AuditoryTestManual* datasets respectively.

Audio-Visual Models: We choose TPAVI (Zhou et al. 2022, 2023) which is the first and state-of-the-art audio-visual semantic segmentation model for comparison with our model. We train it using first-person view and binaural audio as inputs. From Tab. 1, we found that our *SBV* shows strong advantages over TPAVI (Zhou et al. 2022, 2023), not only on the overall performance, but also in out-of-view areas and receives particularly good performance in out-of-view areas. Compared to TPAVI, our SBV improved by 7.7 / 11.4% in the mIoU on out-of-view area, 5.5 / 7.6% on overall performance, and slightly improved by 2.2 / 1.2% on the first-person view, on the *AuditoryTestPseudo* and *AuditoryTestManual* datasets respectively. Fig. 5 shows some segmentation results of TPAVI and our model. We can clearly see that our model segment more objects outside the field of view, and those objects are more defined. This proves that our model can focus not only on in-view objects but also on out-of-view objects. In addition, due to the Omni2Ego dis-

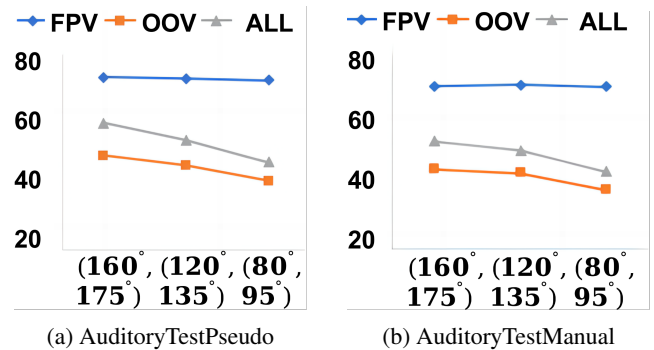(a) AuditoryTestPseudo   (b) AuditoryTestManual

Figure 4: mIoU (%) results for different field of view sizes.

tillation and MRL, our SBV can better reconstruct the information of the out-of-view objects at the feature level. We can see in the fourth row of Fig. 5, our model has a better representation of the shape of the tram at the edge of first-person view compared to TPAVI. Moreover, our model also outperforms better than the 8-channel auditory teacher. We found it is critical to achieve a desired performance using partially missing visual or auditory modality. We achieve satisfactory performance when utilizing both modalities.

Analyses

Impact of Field of View Size We test our model by inputting different sizes of FoV (binocular and monocular) to test the robustness of SBV. We found that the performance is relatively stable and we show the results in Fig. 4. On the whole, we find that the performance in the first-person view (FPV) fluctuates slightly, but as the out-of-view area increases, the overall performance decreases. In a larger FoV (160° width, 175° height), our SBV can achieve better results, that is because SBV can better focus on larger FPV. SBV can still maintain good performance in a small FoV (80° width, 95° height), which should be due to the use of the Omni2Ego distillation and the MRL. Also, our AVFFM makes full use of auditory and visual information, SBV can not only focus on in-view objects but also out-of-view ob-
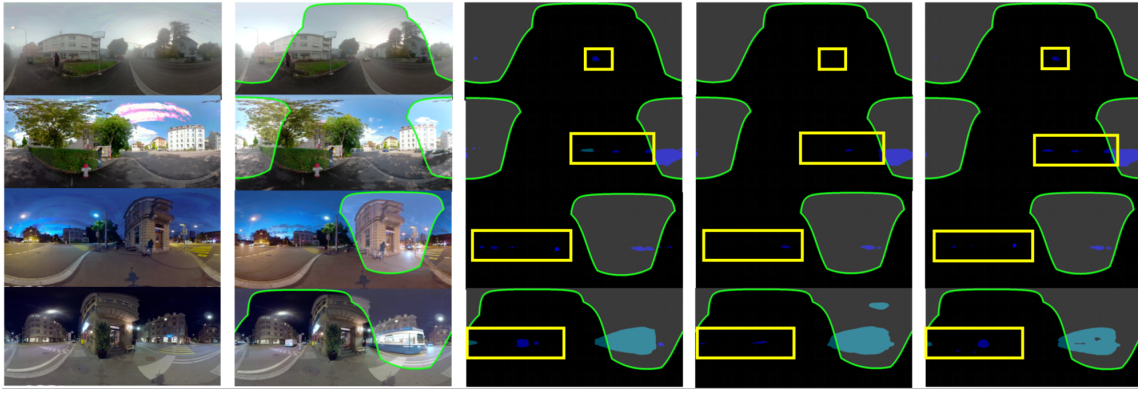
Figure 5: Background, input, ground truth, results of TPAVI (Zhou et al. 2023), ours (SBV) under different weather conditions. Light areas with green lines are first-person views. Yellow boxes mark the out-of-view differences between segmentation results.

Channel	AuditoryTestPseudo			AuditoryTestManual		
	FPV	OOV	All	FPV	OOV	All
Binaural	73.8	46.7	53.0	70.3	43.3	50.1
Mono	73.5	43.2	50.8	70.0	37.8	46.6

Table 2: mIoU (%) results of mono vs. binaural audio input.

Method	AuditoryTestPseudo			AuditoryTestManual		
	FPV	OOV	All	FPV	OOV	All
SBV-full	73.8	46.7	53.0	70.3	43.3	50.1
SBV-v3	72.5	43.5	50.4	69.0	41.0	48.0
SBV-v2	71.7	40.3	48.1	68.7	37.2	45.5
SBV-v1	70.1	38.2	45.9	68.0	34.9	43.7

Table 3: mIoU (%) results of ablation studies.

jects. From another hand, although the FPV is small or any interesting sound-making objects are not included in FPV, it can provide directionality for SBV.

Mono vs. Binaural Many real-world devices at least have a single microphone. We randomly drop an "ear" to do the testing. Results are shown in Tab. 2. We find that performance of SBV does not drop significantly. The FPV mIoU is slightly affected, while the OOV mIoU drops by about 3.5 / 5.5% on *AuditoryTestPseudo* and *AuditoryTestManual* datasets respectively. It means that the segmentation of out-of-view objects are more dependent on the auditory information than in-view objects. That is because binaural audio can provide some position information than mono audio from a cognitive point of view (Blauert 1997; Kendall 1995).

Ablation Studies We conduct ablation studies on Omni2Ego, AVFFM and MRL. Please refer Tab. 3 for results. We first introduce the Omni2Ego and remove this module to verify its effectiveness. We denote it as *SBV-v3*. We find that mIoU decreased slightly (around 1.3% on both test datasets) in the first-person view, indicating that the model learned more shape information of different categories. In addition, the out-of-view mIoU drops about 3 / 2 % mIoU on *AuditoryTestPseudo* / *AuditoryTestManual* test datasets, showing that distillation can indeed help the model reconstruct the missing parts of the modality at the feature level. We then introduce the AVFFM and verify it can indeed help models better focus not only on objects in the first-person view but also on objects outside of the view. We denote it as *SBV-v2* in Tab. 3. We found that the out-of-view mIoU of the model dropped by about 3% on average after removing this module on both test datasets. This shows that AVFFM can help our model get more in-

formation about out-of-view objects from auditory signals. Finally, we introduce the MRL which is expected to help the model to have ability to reconstruct the partially missing modalities. The results of *SBV-v3* in Tab. 3 show that the MRL can help our model improves about 2 % out-of-view and overall mIoU on average on both test datasets.

Conclusion

In this paper, we are the first to introduce and tackle the challenging and novel problem in the field of audio-visual semantic segmentation – Partially Missing Modality issue for multimodal learning. We propose a simple yet efficient framework named Segment Beyond View (SBV) to address this issue. The SBV model leverages Omni2Ego distillation, attention mechanism, and Modality Reconstruction Loss to handle this problem. In the experiments, the proposed model receives promising segmentation accuracy under different evaluation metrics compared to other models. Through extensive analyses, robust performances are achieved with both different sizes of field of view or in mono audio and the effectiveness of each module is further verified. Despite the very exciting out-of-view semantic segmentation result in this paper, the trained model might fails in a completely different scene, e.g. non-urban landscape, with different distribution. This limitation is shared by many similar audio-visual segmentation works (Dai et al. 2022). Examining the generalibility of these models is a promising future direction with significant practical implications. Due to those limitations, we did not explore videos from AR and pedestrian perspectives. As the model matures, it will be beneficial to test the model in controlled street crossing user studies.

Acknowledgments

References

Blauert, J. 1997. *Spatial hearing: the psychophysics of human sound localization.* MIT press.

Bradski, G. 2000. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*.

Campisi, T.; Otković, I. I.; Šurdonja, S.; and Deluka-Tibljaš, A. 2022. Impact of social and technological distraction on pedestrian crossing behaviour: a case study in Enna, Sicily. *Transportation research procedia*, 60: 100–107.

Chakravarthula, P.; D'Souza, J. A.; Tseng, E.; Bartusek, J.; and Heide, F. 2023. Seeing With Sound: Long-range Acoustic Beamforming for Multimodal Scene Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 982–991.

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, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; and Adam, H. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, 801–818.

Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; and Schiele, B. 2016. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3213–3223.

Cortes, C.; Lin, C.; Do, T.; and Chen, H. 2023. An EEG-based Experiment on VR Sickness and Postural Instability While Walking in Virtual Environments. In *2023 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, 94–104. Los Alamitos, CA, USA: IEEE Computer Society.

Dai, D.; Vasudevan, A. B.; Matas, J.; and Van Gool, L. 2022. Binaural soundnet: predicting semantics, depth and motion with binaural sounds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1): 123–136.

Department of Infrastructure, C., Transport; and Government, R. D. A. 2018. Road trauma Australia 2018 statistical summary.

Gan, C.; Zhao, H.; Chen, P.; Cox, D.; and Torralba, A. 2019. Self-supervised moving vehicle tracking with stereo sound. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE.

Gao, R.; Chen, C.; Al-Halah, Z.; Schissler, C.; and Grauman, K. 2020. Visualechoes: Spatial image representation learning through echolocation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, 658–676. Springer.

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

Gruenefeld, U.; Ali, A. E.; Boll, S.; and Heuten, W. 2018. Beyond Halo and Wedge: visualizing out-of-view objects on head-mounted virtual and augmented reality devices. In *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services*, number Article 40 in MobileHCI '18, 1–11. New York, NY, USA: Association for Computing Machinery.

Gruenefeld, U.; Ennenga, D.; Ali, A. E.; Heuten, W.; and Boll, S. 2017. EyeSee360: Designing a Visualization Technique for out-of-View Objects in Head-Mounted Augmented Reality. SUI '17, 109–118. New York, NY, USA: Association for Computing Machinery. ISBN 9781450354868.

Gustafson, S. G.; and Irani, P. P. 2007. Comparing visualizations for tracking off-screen moving targets. In *CHI'07 Extended Abstracts on Human Factors in Computing Systems*, 2399–2404.

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.

Holländer, K.; Krüger, A.; and Butz, A. 2020. Save the Smombies: App-Assisted Street Crossing. In *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services*, number Article 22 in MobileHCI '20, 1–11. New York, NY, USA: Association for Computing Machinery.

Hu, D.; Qian, R.; Jiang, M.; Tan, X.; Wen, S.; Ding, E.; Lin, W.; and Dou, D. 2020. Discriminative sounding objects localization via self-supervised audiovisual matching. *Advances in Neural Information Processing Systems*, 33: 10077–10087.

Jung, J.; Lee, H.; Choi, J.; Nanda, A.; Gruenefeld, U.; Stratmann, T.; and Heuten, W. 2018. Ensuring Safety in Augmented Reality from Trade-off Between Immersion and Situation Awareness. In *2018 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, 70–79.

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.

Kang, H.; Lee, G.; and Han, J. 2019. Obstacle Detection and Alert System for Smartphone AR Users. In *Proceedings of the 25th ACM Symposium on Virtual Reality Software and Technology*, VRST '19. New York, NY, USA: Association for Computing Machinery. ISBN 9781450370011.

Kendall, G. S. 1995. A 3-D sound primer: directional hearing and stereo reproduction. *Computer music journal*, 19(4): 23–46.

Kim, J. U.; and Kim, S. T. 2023. Towards Robust Audio-Based Vehicle Detection Via Importance-Aware Audio-Visual Learning. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

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

Li, Y.; Liu, H.; and Tang, H. 2022. Multi-modal perception attention network with self-supervised learning for audio-visual speaker tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 1456–1463.

Liu, J.; Ju, C.; Xie, W.; and Zhang, Y. 2022. Exploiting transformation invariance and equivariance for self-supervised sound localisation. In *Proceedings of the 30th ACM International Conference on Multimedia*, 3742–3753.

Ma, M.; Ren, J.; Zhao, L.; Testuggine, D.; and Peng, X. 2022. Are multimodal transformers robust to missing modality? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18177–18186.

Ma, M.; Ren, J.; Zhao, L.; Tulyakov, S.; Wu, C.; and Peng, X. 2021. Smil: Multimodal learning with severely missing modality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 2302–2310.

Manori, G.; Singh, G.; Dushad, M.; Vinod, M.; Choudekar, P.; Tayal, V. K.; et al. 2018. Acoustic Detection of Rear Approaching Vehicles for Cyclists. In *2018 International Conference on Power Energy, Environment and Intelligent Control (PEEIC)*, 354–357. IEEE.

Miech, A.; Laptev, I.; and Sivic, J. 2018. Learning a text-video embedding from incomplete and heterogeneous data. *arXiv preprint arXiv:1804.02516*.

Mizumachi, M.; Kaminuma, A.; Ono, N.; and Ando, S. 2014. Robust sensing of approaching vehicles relying on acoustic cues. *Sensors*, 14(6): 9546–9561.

Recasens, A.; Lin, J.; Carreira, J.; Jaegle, D.; Wang, L.; Alayrac, J.-b.; Luc, P.; Miech, A.; Smaira, L.; Hemsley, R.; et al. 2023. Zorro: the masked multimodal transformer. *arXiv preprint arXiv:2301.09595*.

Rovetta, S.; Mnasri, Z.; and Masulli, F. 2020. Detection of hazardous road events from audio streams: An ensemble outlier detection approach. In *2020 IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS)*, 1–6. IEEE.

Samsonov, V. 2017. Improved background subtraction algorithm - OpenCV.

Serino, M.; Cordrey, K.; McLaughlin, L.; and Milanaik, R. L. 2016. Pokémon Go and augmented virtual reality games: a cautionary commentary for parents and pediatricians. *Curr. Opin. Pediatr.*, 28(5): 673–677.

Shvetsova, N.; Chen, B.; Rouditchenko, A.; Thomas, S.; Kingsbury, B.; Feris, R. S.; Harwath, D.; Glass, J.; and Kuehne, H. 2022. Everything at once-multi-modal fusion transformer for video retrieval. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 20020–20029.

Smith, L. N.; and Topin, N. 2019. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial intelligence and machine learning for multi-domain operations applications*, volume 11006, 369–386. SPIE.

Tian, Y.; Shi, J.; Li, B.; Duan, Z.; and Xu, C. 2018. Audio-visual event localization in unconstrained videos. In *Proceedings of the European conference on computer vision (ECCV)*, 247–263.

Tong, Y.; Jia, B.; and Bao, S. 2021. An Augmented Warning System for Pedestrians: User Interface Design and Algorithm Development. *Applied Sciences*, 11: 7197.

Valverde, F. R.; Hurtado, J. V.; and Valada, A. 2021. There is more than meets the eye: Self-supervised multi-object detection and tracking with sound by distilling multimodal knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11612–11621.

Vasudevan, A. B.; Dai, D.; and Van Gool, L. 2020. Semantic object prediction and spatial sound super-resolution with binaural sounds. In *European conference on computer vision*, 638–655. Springer.

Wandell, B. A. 1995. *Foundations of vision*. sinauer Associates.

Wang, H.; Chen, Y.; Ma, C.; Avery, J.; Hull, L.; and Carneiro, G. 2023. Multi-Modal Learning With Missing Modality via Shared-Specific Feature Modelling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15878–15887.

Wang, T.; Cardone, G.; Corradi, A.; Torresani, L.; and Campbell, A. T. 2012. WalkSafe: a pedestrian safety app for mobile phone users who walk and talk while crossing roads. In *Proceedings of the Twelfth Workshop on Mobile Computing Systems & Applications*, number Article 5 in HotMobile '12, 1–6. New York, NY, USA: Association for Computing Machinery.

Wu, R.; and Chen, H. 2023. The Effect of Visual and Auditory Modality Mismatching between Distraction and Warning on Pedestrian Street Crossing Behavior. In *2023 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, 1045–1054. Los Alamitos, CA, USA: IEEE Computer Society.

Wu, Y.; Zhu, L.; Yan, Y.; and Yang, Y. 2019. Dual attention matching for audio-visual event localization. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6292–6300.

Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J. M.; and Luo, P. 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34: 12077–12090.

Zhou, J.; Shen, X.; Wang, J.; Zhang, J.; Sun, W.; Zhang, J.; Birchfield, S.; Guo, D.; Kong, L.; Wang, M.; and Zhong, Y. 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 *European Conference on Computer Vision*.

Zürn, J.; and Burgard, W. 2022. Self-supervised moving vehicle detection from audio-visual cues. *IEEE Robotics and Automation Letters*, 7(3): 7415–7422.