Panoramic Video Salient Object Detection with Ambisonic Audio Guidance

Xiang Li[1]*, Haoyuan Cao[2], Shijie Zhao[3], Junlin Li[2], Li Zhang[2], Bhiksha Raj[1,4]

[1] Carnegie Mellon University, PA, USA.
[2] Bytedance Inc., San Diego, CA, USA.
[3] Bytedance Inc., Shenzhen, China.
[4] Mohammed bin Zayed University of AI, Abu Dhabi, UAE.
{xl6, bhiksha}@andrew.cmu.edu, {haoyuan.cao, zhaoshijie.0526, lijunlin.li, lizhang.idm}@bytedance.com

Abstract

Video salient object detection (VSOD), as a fundamental computer vision problem, has been extensively discussed in the last decade. However, all existing works focus on addressing the VSOD problem in 2D scenarios. With the rapid development of VR devices, panoramic videos have been a promising alternative to 2D videos to provide immersive feelings of the real world. In this paper, we aim to tackle the video salient object detection problem for panoramic videos, with their corresponding ambisonic audios. A multimodal fusion module equipped with two pseudo-siamese audio-visual context fusion (ACF) blocks is proposed to effectively conduct audio-visual interaction. The ACF block equipped with spherical positional encoding enables the fusion in the 3D context to capture the spatial correspondence between pixels and sound sources from the equirectangular frames and ambisonic audios. Experimental results verify the effectiveness of our proposed components and demonstrate that our method achieves state-of-the-art performance on the ASOD60K dataset.

Introduction

Video salient object detection (VSOD) aims to find the most visually distinctive objects in a video. VSOD for 2D videos has been attracting considerable attention (Su et al. 2022; Liu et al. 2021; Ren et al. 2021b) due to its wide applications in real-world scenarios, such as video editing and video compression. While, for panoramic videos which have a very different format and viewing environment, how to effectively detect salient objects is still an open problem. Since human attention is usually influenced by acoustic signatures that are naturally synchronized with visual objects in audio-bearing video recordings, some VSOD methods for 2D videos (Tsiami, Koutras, and Maragos 2020; Cheng et al. 2021) introduce acoustic modality to facilitate the saliency discrimination. Different from mono or binaural audio used in 2D videos, ambisonic audio is utilized to create immersive feelings of the real world in panoramic videos. In this work, we focus on how to detect the salient objects in panoramic videos with their corresponding ambisonic audios. To the best of our knowledge, we are the first to tackle the VSOD problem for panoramic scenarios.

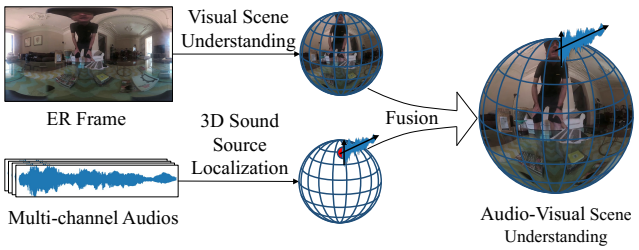

Figure 1: We study the problem of how to utilize the ambisonic audio to facilitate the panoramic video salient object detection. Since ambisonic audio contains rich spatial information, we can directly localize the sound sources and then fuse them with visual cues. On the other hand, since the ER frame has distortions and cannot reflect the 3D position of pixels, projecting the pixels back to 3D space is essential for effective multimodal fusion.

The panoramic video contains omnidirectional contexts and represents each pixel on a 3D sphere. Unlike 2D videos that display with a stable viewpoint and fixed environmental audios, panoramic videos are typically supported by VR headsets which provide a head direction-adaptive field of view (FOV) and ambisonic audios. With multi-channel audio recordings, ambisonic audios encode the 3D location of the sound source and the VR headset can further adjust the original audio to provide a real sense of sound source to the user given the movement of the head. Previous works (Tsiami, Koutras, and Maragos 2020; Cheng et al. 2021) have demonstrated the audio shows non-trivial influence on human attention in 2D videos. For VSOD in panoramic videos, as shown in Figure 1, since the ambisonic audio can directly reflect the 3D sound source location, we consider it should perform a more important role compared to 2D scenarios.

On the other hand, previous panoramic video processing approaches cannot preserve the spatial relationship of panoramic videos. Due to the unique format of panoramic videos, it is difficult to store or transmit the raw video using current video coding techniques. Therefore, equirectangular (ER) projection is commonly leveraged to transform the panoramic video into a regular 2D format. However, the ER projection will involve not only a separation along the longitude of the sphere but also distortions in the polar area,

*This work was done when Xiang Li was an intern at Bytedance.

which severely barrier the effective VSOD. Previous methods address the polar distortions by introducing a cube projection (Cheng et al. 2018) which projects the sphere to a cube and expands each face to ease the distortions. However, the padded cube map severely destroys the spatial relationship between each face which obstacles the global understanding of the frame.

In this paper, we purpose a framework for audio-visual salient object detection in panoramic scenarios. Considering the rich spatial and semantic information encoded in the ambisonic audios, we first use a pretrained acoustic encoder to extract location and semantic embeddings of 3D sound sources, and then introduce audio-visual context fusion (ACF) blocks to enhance the visual features by acoustic cues. Ideally, the 3D location of sound sources should be utilized as ground-truth to finetune the acoustic network while it is very difficult to obtain for common panoramic videos. To tackle this problem, inspired by label-guided distillation (Zhang et al. 2022), the ground-truth label of salient object is employed in the multimodal fusion by enforcing consistency between teacher (equipped with label) and student branch. By learning better audio-visual correspondence, we can transfer the spatial information encoded in the visual objects to supervise acoustic modality. In particular, to reflect the true 3D location of objects and mitigate the influence of distortions in ER frames, we map the 2D coordinates of pixels back to the 3D sphere and encode them in the positional encoding during the fusion. In this way, the model can capture the true spatial position of each pixel. Our contributions are summarized as:

- We propose an audio-visual video salient object detection framework for panoramic scenarios. To the best of our knowledge, we are the first to tackle this problem.
- We introduce a label-guided audio-visual fusion module to effectively utilize the rich spatial and semantic information encoded in the ambisonic audio recordings synchronized with panoramic videos.
- Our model achieves state-of-the-art results on the ASOD60K dataset. Extensive experiments are conducted to illustrate the effectiveness of our method.

Related Works

Video Salient Object Detection

VSOD aims to find the most visually salient objects in the video. Conventional methods usually leverage color contrast (Achanta et al. 2009), motion prior (Zhang, Yu, and Crandall 2019), background prior (Yang et al. 2013) and center prior (Bertasius et al. 2017) to distinguish the salient regions. However, most of those methods are limited by the low representative ability of hand-crafted features. Recent methods leverage deep-learning approaches to tackle the VSOD problem. To utilize the temporal information, FCNS (Wang, Shen, and Shao 2017) first leverages FCN for static saliency prediction and then post-process them by another dynamic FCN. Similarly, DSRFCN3D (Le and Sugimoto 2017) introduces 3D convolution for temporal aggregation. Optical flow reflecting the motion also serves as a strong cue for VSOD task in (Li et al. 2018, 2019). Recently, more advanced structures are leveraged to better understand the spatio-temporal correspondence. For example, ConvLSTM (Li et al. 2018; Fan et al. 2019) is adopted to construct long-term temporal relation. With the strong ability of transformer (Vaswani et al. 2017) to model the complex relationship, it has achieved promising result in VSOD (Liu et al. 2021; Ren et al. 2021b).

Panoramic Saliency Detection

Saliency detection aims to predict the region of human eye fixation in the video. For image-level saliency prediction (Cheng et al. 2018; Suzuki and Yamanaka 2018),(Cheng et al. 2018) propose a cube padding operation to project the panoramic frame to a cube with fewer distortions on each face. (Zhu, Zhai, and Min 2018) first map the ER frame to a sphere then predict the saliency from each viewport. For video-level saliency prediction (Cheng et al. 2018; Nguyen, Yan, and Nahrstedt 2018), Nguyen et al. (Nguyen, Yan, and Nahrstedt 2018; Zhang et al. 2018) proposes a method that leverages CNN and LSTM for saliency prediction. In (Zhang et al. 2018), spherical CNN is introduced to directly handle panoramic videos. In addition, audio is introduced in panoramic saliency prediction (Chao et al. 2020) given its strong ability to influence human attention.

Video Object Segmentation

Video object segmentation (VOS) can be categorized as unsupervised (Wang et al. 2019; Ren et al. 2021a), semi-supervised (Wang et al. 2021) and referring (Wu et al. 2022; Li et al. 2022c) VOS. The most relevant type to this work is the unsupervised VOS (UVOS) which aims to segment primary object regions from the background in videos. Early methods tackle the UVOS problem by object proposal (Kim and Hwang 2002), temporal trajectory (Fragkiadaki, Zhang, and Shi 2012) and saliency prior (Wang et al. 2015; Wang, Shen, and Porikli 2015). More recently, deep learning-based methods are proposed for modeling the spatio-temporal information. MATNet (Zhou et al. 2020) uses a motion-attentive transition to model motion cues and spatio-temporal representation. RTN (Ren et al. 2021a) leverages long-range intra-frame contrast, temporal coherence, and motion-appearance similarity to enhance the appearance feature representation. In addition to VOS, video instance segmentation (Li et al. 2022a,b,d) is also relevant to this work. Recently, some works extend the video segmentation task to multimodal by considering audio (Zhou et al. 2022) or signals (Zhao et al. 2022a; Huang et al. 2021b,a).

Method

Overview

Given a video clip $V = \{I_t\}_{t=1}^{T}$ of T frames and its corresponding multi-channel audio recordings $A = \{H_i\}_{i=1}^{4}$, we predict the salient object $\{M_t\}_{t=1}^{T}$ effectively with our method. The method overview is illustrated in Figure 2. The pipeline can be boiled down into three parts: acoustic and visual encoders, a label-guided multimodal fusion module, and decoders. We first leverage a visual and an acoustic encoder to extract visual $\{f_t\}_{t=1}^{T}$ and acoustic features g^{sem}, g^{loc}. The acoustic features contains a semantic embedding g^{sem}

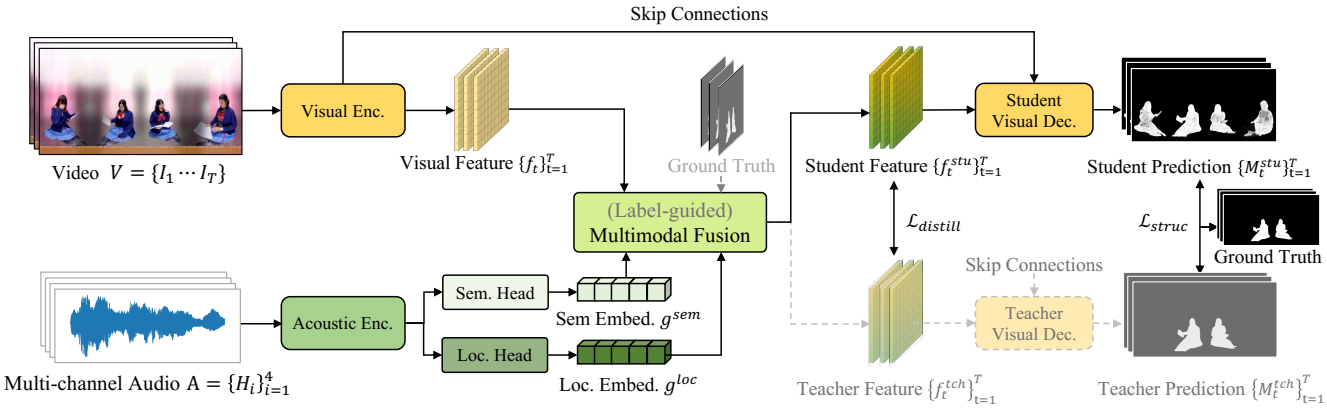

Figure 2: Pipeline Overview. We use separate encoders to extract multimodal features. For a video clip $V = \{I_1 \cdots I_T\}$, a visual encoder is utilized to extract visual feature $\{f_t\}_{t=1}^T$. For audio input $A = \{H_i\}_{i=1}^4$, a two-brunch acoustic encoder is employed to extract the semantic embedding g^{sem} and location embedding g^{loc}. After that, a label-guided multimodal fusion module is introduced to effectively fuse the multimodal features, which outputs a student feature $\{f_t^{stu}\}_{t=1}^T$ and a teacher feature $\{f_t^{tch}\}_{t=1}^T$. Two decoders are leveraged to decode the final predictions $\{M_t^{stu}\}_{t=1}^T$ and $\{M_t^{tch}\}_{t=1}^T$ from compacted features $\{f_t^{stu}\}_{t=1}^T$ and $\{f_t^{tch}\}_{t=1}^T$ respectively. In particular, to enhance multimodal communication, a distillation loss $\mathcal{L}_{ditill}$ is adopted between student feature $\{f_t^{stu}\}_{t=1}^T$ and (label-guided) teacher feature $\{f_t^{tch}\}_{t=1}^T$. A structure loss $\mathcal{L}_{struc}$ is utilized as the objective. The **gray** color indicates components that are only used during training.

and a location embedding g^{loc} which encodes the category and 3D location of sound sources respectively. After that, a multimodal fusion module is utilized to enable audio-visual interaction. Specifically, the multimodal fusion module contains two pseudo-siamese blocks - a student block that fuses audio-visual information using multimodal attention and a teacher block that shares the same structure of the student block while taking an additional ground truth as input to guide the fusion. The output student features $\{f_t^{stu}\}_{t=1}^T$ and (label-guided) teacher features $\{f_t^{tch}\}_{t=1}^T$ are sent to visual decoders equipped with skip connections to generate the final salient object predictions $\{M_t^{stu}\}_{t=1}^T$ and $\{M_t^{tch}\}_{t=1}^T$. A distillation loss between student feature $\{f_t^{stu}\}_{t=1}^T$ and (label-guided) teacher features $\{f_t^{tch}\}_{t=1}^T$ is adopted to help the multimodal interaction and a structure loss between prediction $\hat{M}_t$ and ground truth M_t is used as the task objective for salient object detection.

Label-guided Multimodal Fusion

The label-guided multimodal fusion module contains two pseudo-siamese audio-visual context fusion (ACF) blocks equipped with spherical positional encoding to align the correspondence between 3D sound sources and pixels. The output of student and teacher block are student feature $\{f_t^{stu}\}_{t=1}^T$ and teacher feature $\{f_t^{tch}\}_{t=1}^T$ respectively.

Spherical positional encoding. ER frame is a commonly used format to transmit and store panoramic videos (Cai et al. 2022). However, as shown in Figure 3, the ER frame suffers from severe distortions in the polar regions. To tackle this problem, we adopt the popular position-agnostic attention mechanism (Vaswani et al. 2017; Zhao, Guo, and Lu 2022; Zhao et al. 2022b) and propose a spherical positional encoding to compensate for the distortion in the ER

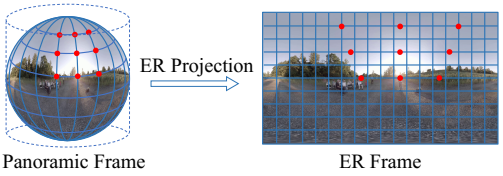

Figure 3: Illustration of ER Projection. Severe distortions can be observed in the polar areas.

frame. Different from regular positional encoding (Vaswani et al. 2017), spherical positional encoding first re-projects the plane coordinates of each pixel back to the 3D sphere and generates positional encoding based on the 3D coordinates. In this way, each pixel can reflect its true 3D position thus avoiding the severe distortion in the polar region and inevitable separation along a longitude in the ER frame. In particular, the spherical positional encoding can be computed as $\mathrm{PE}(\mathrm{pos}_{\mathrm{3D}}, 2i) = \sin(\mathrm{pos}_{\mathrm{3D}}/10000^{2i/\frac{d}{3}})$ and $\mathrm{PE}(\mathrm{pos}_{\mathrm{3D}}, 2i+1) = \cos(\mathrm{pos}_{\mathrm{3D}}/10000^{2i/\frac{d}{3}})$ where $\mathrm{pos}_{\mathrm{3D}}$ can be x, y, z coordinates and i is the dimension. The 2D-to-3D transformation can be computed as

$$x = sin\frac{v}{R}cos\frac{u}{R},\ y = sin\frac{v}{R}sin\frac{u}{R},\ z = cos\frac{v}{R} \quad (1)$$

where (u, v) and (x, y, z) are the 2D and 3D coordinate of each pixel respectively. $R = \frac{W}{2\pi}$ where W is the width of the frame. Spherical positional encoding (SPE) is employed to encode spatial information for visual representation during cross-modal attention.

Student block. As shown in Figure 4 (without the gray parts), to fuse the rich information encoded in visual and acoustic features, we utilize multimodal attention to enable audio-visual context interaction. We first concatenate

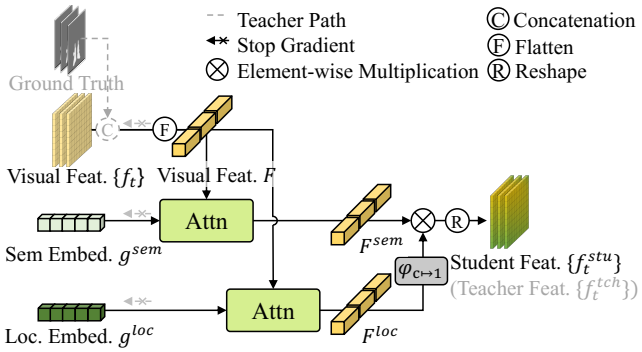

Figure 4: Illustration of ACF block used in the (label-guided) multimodal fusion process. Student block: The visual feature $\{f_t\}_{t=1}^T$ is first flattened and added with positional encoding then conducts multimodal attention with acoustic semantic embedding g^{sem} and location embedding g^{loc} respectively. After that, we form a pixel-wise weighting from fused feature F^{loc} to modulate F^{sem}. The output of student block is denoted as $\{f_t^{stu}\}_{t=1}^T$. Teacher block: Different from student block, we concatenate ground-truth mask with the visual feature $\{f_t\}_{t=1}^T$ to help the network learn better representation in the multimodal fusion. Since teacher block is only employed during training, we truncate the gradients before the multimodal fusion in the teacher block. The output of teacher block is denoted as $\{f_t^{tch}\}_{t=1}^T$.

and then flatten the visual feature $\{f_t\}_{t=1}^T$ to form $F = flatten(f_1 \oplus \cdots \oplus f_T) \in \mathbb{R}^{C\times THW}$. After that, spherical and regular positional encodings (Vaswani et al. 2017) are added to visual feature F, and acoustic feature g^{sem} and g^{loc}, respectively, to help the network capture spatial information. The multimodal attention can be computed by

$$h^{aud} = \mathrm{LN}(\mathrm{MCA}(F, g^{aud}) + F) \tag{2}$$

$$F^{aud} = \mathrm{LN}(\mathrm{FFN}(h^{aud}) + h^{aud}) \tag{3}$$

where MCA, FFN and LN are multi-head cross-attention (Ye et al. 2019), feed-forward network and layer-normalization respectively. In particular, we generate the query Q from F and key K, value V from g^{aud} by linear projections. The $\mathrm{MCA}(F, g^{aud})$ can be computed by $\mathrm{Softmax}(\frac{K^{\mathrm{T}}Q}{\sqrt{d}})V$, where d is the dimension of query Q. The student feature F^{stu} can be computed by

$$F^{stu} = F^{sem} \odot \varphi_{C\mapsto 1}(F^{loc}) \tag{4}$$

where $\varphi_{C\mapsto 1}$ denotes a convolution to reduce channel from C to 1 and $\odot$ denotes element-wise multiplication. The final output is $\{f_t^{stu}\}_{t=1}^T = \mathrm{Reshape}(F^{stu})$.

Teacher block. The audio encoder is pretrained on a 3D sound source localization dataset (Guizzo et al. 2021) while it is difficult to obtain a 3D location of vocal objects in panoramic videos during main training. Inspired by previous work (Zhang et al. 2022), we build a pseudo-siamese teacher block to help the network capture accurate spatial information by introducing ground-truth annotation to the multimodal fusion. As shown in Figure 4 (with the gray parts), the additional ground truth is downsampled and concatenated with visual feature $\{f_t\}_{t=1}^T$ in the teacher block before conducting the multimodal interaction. Similar to student block, the final output of teacher block is denoted as $\{f_t^{tch}\}_{t=1}^T$. Note that the teacher block does not share weights with the student block and the gradient of teacher block is truncated to avoid influencing the encoder.

Encoder

We use separate encoders to extract visual and acoustic features.

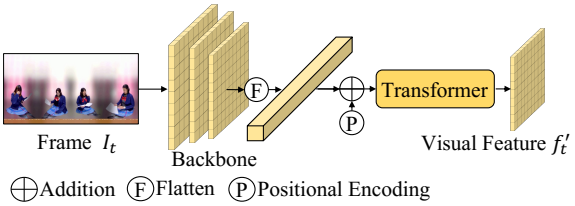

Figure 5: Single-frame visual feature extraction. The ER frame I_t is first processed by a backbone and then the extracted features are fed to a transformer to capture spatial information. The single-frame feature is denoted as f_t'.

Visual encoder. We first project the panoramic video to 2D frames $\{I_1 \cdots I_T\}$ using ER projection and then feed them to the backbone. As shown in Figure 5, a transformer encoder on top of the ResNet-50 (He et al. 2016) is adopted to mitigate the severe distortion in the ER frame. In addition, a temporal non-local block as (Yan et al. 2019) is also leveraged to enable the temporal interaction on the extracted features $\{f_t'\}_{t=1}^T$ from the backbone. We denote the features after temporal aggregation as $\{f_t\}_{t=1}^T$ where $f_t \in \mathbb{R}^{C\times H\times W}$.

Acoustic encoder. Multi-channel audio contains the 3D location and category information of the sound source which have a great impact on the choosing of the salient objects. To extract the acoustic features, we leverage three CNN layers followed by two bi-directional GRU layers as our acoustic encoder and three linear layers for both semantic and location head (detailed structure available in supplementary) (Adavanne et al. 2018). Since it is difficult to obtain the real-world sound source location in panoramic videos, we first pretrain the acoustic encoder on a 3D sound source localization and sound event classification dataset, L3DAS (Guizzo et al. 2021). We remove final linear layer in each head to form the semantic embedding $g^{sem} \in \mathbb{R}^{C\times L}$ and location embedding $g^{loc} \in \mathbb{R}^{C\times L}$.

Decoder

We adopt the same structure for decoding $\{f_t^{stu}\}_{t=1}^T$ and $\{f_t^{tch}\}_{t=1}^T$. For decoding the salient object prediction, we follow the FPN structure (Lin et al. 2017) to fuse the low-level features. Let the output salient object prediction be $\{M_t^{stu}\}_{t=1}^T \in \mathbb{R}^{H_o\times W_o}$ and $\{M_t^{tch}\}_{t=1}^T \in \mathbb{R}^{H_o\times W_o}$ for student and teacher branch respectively, where H_o and W_o are the height and width of the output.

Method	Miscellanea (Test1)				Music (Test2)				Speaking (Test3)				ASOD60K-Test All			
	$F_\beta \uparrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$\mathcal{M} \downarrow$	$F_\beta \uparrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$\mathcal{M} \downarrow$	$F_\beta \uparrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$\mathcal{M} \downarrow$	$F_\beta \uparrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$\mathcal{M} \downarrow$
Image-Level Methods																
CPD-R	.248	.654	.645	.035	.272	.608	.632	.018	.228	.588	.657	.026	.243	.609	.648	.026
SCRN	.250	.665	.615	.046	.341	.683	.664	.023	.276	.636	.642	.034	.286	.655	.641	.034
F3Net	.257	.655	.629	.040	.358	.662	.749	.021	.308	.626	.692	.027	.310	.642	.691	.029
MINet	.238	.650	.625	.050	.380	.670	.716	.020	.261	.590	.635	.053	.286	.624	.652	.044
LDF	.280	.663	.626	.044	.389	.671	.753	.023	.309	.625	.711	.037	.322	.645	.701	.035
CSFR2	.238	.652	**.642**	.033	.347	.665	.693	.018	.285	.636	.700	.026	.290	.646	.684	.026
GateNet	.285	.677	.651	.044	.290	.673	.616	.018	.260	.633	.638	.034	.273	.653	.636	.033
Video-Level Methods																
COSNet	.147	.610	.553	**.031**	.220	.557	.541	.016	.176	.572	.570	**.023**	.181	.582	.559	**.023**
RCRNet	.272	.661	.640	.034	.403	**.695**	.738	.019	.282	.632	.687	.030	.310	.654	.688	.029
PCSA	.123	.604	.574	.034	.310	.657	.645	.022	.150	.571	.534	.026	.184	.600	.570	.027
3DC-Seg	.300	.668	.618	.062	.326	.635	.632	.046	.289	.629	.592	.056	.300	.640	.608	.055
RTNet	.240	.622	.634	.038	.365	.638	.766	.020	.194	.555	.668	.028	.247	.591	.683	.025
Ours	**.355**	**.714**	.617	.037	**.483**	.682	**.834**	**.016**	**.382**	**.658**	**.742**	.024	**.404**	**.678**	**.732**	.026

Table 1: Comparison to state-of-the-art salient object detection methods on ASOD60K dataset. ↑ means larger is better and ↓ means smaller is better. Bold means the state-of-the-art performance.

Loss Function

The overall objective of our proposed method is composed of structure losses for student and teacher branch $\mathcal{L}_{struc}^{stu}$, $\mathcal{L}_{struc}^{tch}$ and a distillation loss $\mathcal{L}_{distill}$

$$\mathcal{L} = \mathcal{L}_{struc}^{stu} + \mathcal{L}_{struc}^{tch} + \lambda_{distill}\mathcal{L}_{distill} \tag{5}$$

where $\lambda_{distill}$ is a scalar to balance the losses.

Structure loss. Following previous method (Chen et al. 2022), we leverage a combination of binary cross entropy loss and Dice loss (Milletari, Navab, and Ahmadi 2016) as the objective for salient object detection.

$$\mathcal{L}_{struc} = \sum_{t=1}^{T} \mathcal{L}_{bce}(M_t, \hat{M}_t) + \lambda_{dice} \sum_{t=1}^{T} \mathcal{L}_{dice}(M_t, \hat{M}_t) \tag{6}$$

where M_t and $\hat{M}_t$ are predicted and ground-truth salient maps respectively. λ_{dice} is a scalar. M_t can be M_t^{stu} and M_t^{tch} for computing $\mathcal{L}_{struc}^{stu}$ and $\mathcal{L}_{struc}^{tch}$.

Distillation loss. To help the student block learn the audio-visual correspondence, we enforce a consistency between f_t^{tch} and f_t^{stu}. A MSE loss is utilized as the constraint

$$\mathcal{L}_{distill} = \sum_{t=1}^{T} \mathcal{L}_{MSE}(f_t^{tch}, f_t^{stu}) \tag{7}$$

Inference

Since the purpose of introducing teacher block is to facilitate network learn accurate audio-visual correspondence during training, we disable the teacher block and only keep the student block during inference.

Experiment

Dataset and Metrics

Dataset. We conduct experiments on the ASOD60K dataset (Zhang, Chao, and Zhang 2021) which is an audio-induced salient object detection benchmark for panoramic videos. There are 62,455 frames with 10,465 instance-level ground truths in the dataset. In particular, each video corresponds to a 4-channel ambisonic audio recording. The ground-truth salient objects are determined by the eye fixation of 40 participants who viewed the video with HTC Vive HMD headset. The test set of ASOD60K contains three subsets split by sound event classes - miscellanea, music, and speaking.

Metrics. To evaluate the performance of video salient object detection, we employ the adaptive F-Measure F_β (Achanta et al. 2009), adaptive E-Measure E_ϕ (Fan et al. 2018), S-Measure S_α (Fan et al. 2017) and Mean Absolute Error (MAE) $\mathcal{M}$ (Borji et al. 2015).

Implementation Details

Following the benchmark setting(Zhang, Chao, and Zhang 2021), our model is first pre-trained on DUST dataset (Wang et al. 2017) and then finetuned on ASOD60K (Zhang, Chao, and Zhang 2021). The model is trained for 20 epochs with a learning rate of 1e-4. We adopt a batchsize of 2 and an AdamW (Loshchilov and Hutter 2017) optimizer with weight decay 0. All images are cropped to have the longest side of 832 pixels and the shortest side of 416 pixels during training and evaluation. The window size is set to 3. The $\lambda_{distill}$ is set to 5.0 and λ_{dice} is set to 1 if no specification. We leverage a 3-layer transformer encoder (Dosovitskiy et al. 2020) on top of the ResNet-50 (He et al. 2016) to extract visual features. We leverage an augmented SELDNet (Adavanne et al. 2018) as our acoustic encoder. Our method is implemented with PyTorch.

Main Results

In this section, we compare our method with previous state-of-the-art methods, including CPD-R (Wu, Su, and Huang 2019a), MINet (Pang et al. 2020), SCRN (Wu, Su, and Huang

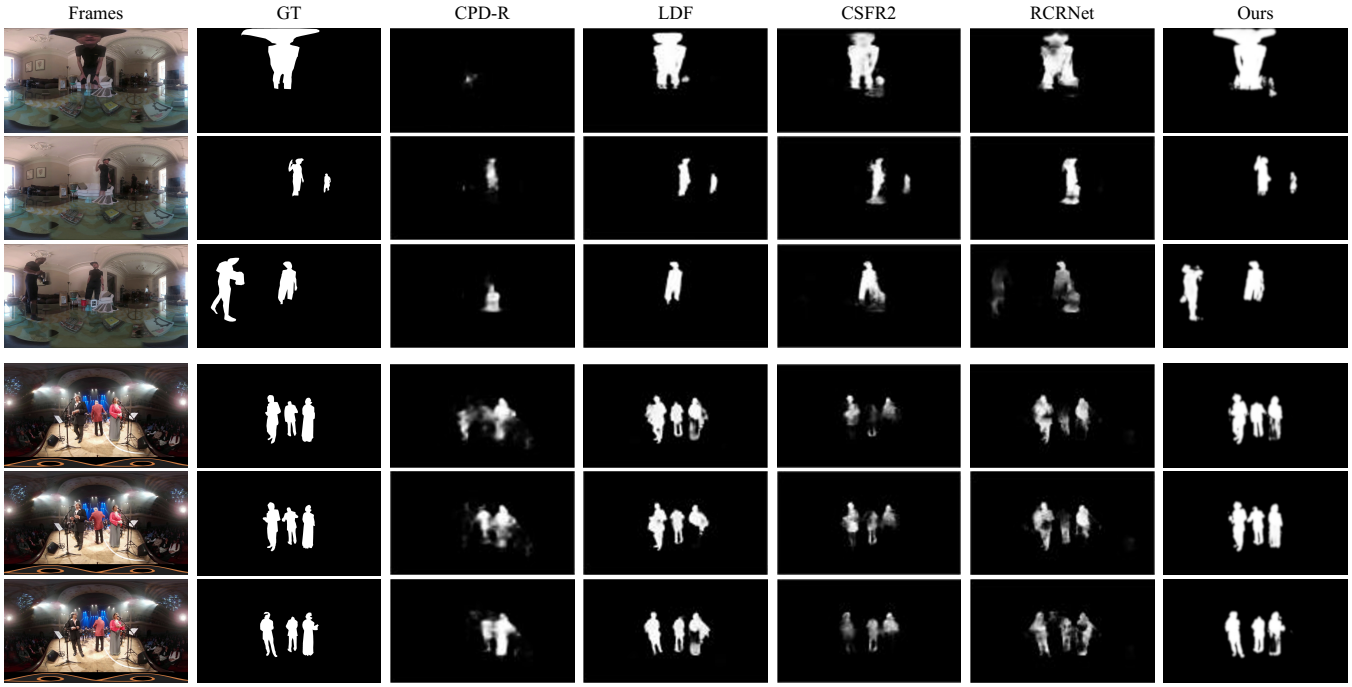

Figure 6: Qualitative comparison to state-of-the-art VSOD methods on ASOD60K dataset.

2019b), F3Net (Wei, Wang, and Huang 2020), LDF (Wei et al. 2020), CSFR2 (Gao et al. 2020), GateNet (Zhao et al. 2020), COSNet (Lu et al. 2019), RCRNet (Yan et al. 2019), PCSA (Gu et al. 2020), 3DC-Seg (Mahadevan et al. 2020) and RTNet (Ren et al. 2021a) on ASOD60K dataset.

Quantitative results. We compare our method with state-of-the-art methods on the ASOD60K dataset in Table 1. In general, our method achieves the best result of 0.404 F_β, 0.678 S_α, 0.732 E_ϕ and 0.026 $\mathcal{M}$ on the ASOD60K test set. For each sound event split, all metrics of our method eclipse other methods on both Music and Speaking splits. While the 0.617 E_ϕ of our method on the Miscellanea split is slightly lower than the 0.644 E_ϕ of RCRNet (Yan et al. 2019). Two reasons maybe account for the inferior performance of our method. First, the audio recordings in Miscellanea split contain background music which barriers our model to accurately locating the sound sources. Second, there are several unseen sound event classes in the Miscellanea test split. In this way, it is difficult for the model to construct correct multimodal correspondence between videos and audios with unseen classes. In the music and speaking scenario where sound sources can be easily localized, our method achieves obvious improvement against previous methods.

Figure 7: Visualization of sound source localization heatmap.

Multimodal Fusion	ASOD60K-Test All			
	$F_\beta \uparrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$\mathcal{M} \downarrow$
None	.396	.660	.714	.037
+ACF (Concat)	.385	.667	.697	.027
+ACF (MM Attn)	.397	.670	.722	**.026**
+ACF (MM Attn)+SPE	**.404**	**.678**	**.732**	**.026**

Table 2: Impact of different multimodal fusion methods. The content in the bracket indicates different fusion methods in ACF block. Concat: Concatenate, MM Attn: multimodal attention, SPE: spherical positional encoding.

Qualitative results. We present our qualitative result in Figure 6 and compare it against previous 2D methods on ASOD60K dataset. The result indicates that previous methods fail to detect the correct salient objects. In contrast, our method shows great accuracy and robustness even in very challenging scenarios, e.g., with severe distortions in the polar area as shown in the first line in Figure 6. This implies that our network equipped with ACF block and SPE generates more accurate results than simply adopting previous 2D methods on the panoramic scenario. We visualize the audio-guided location heatmap $\varphi_{C\mapsto 1}(f_t^{loc})$ in Figure 7. We notice that the audio-guided location heatmap $\varphi_{C\mapsto 1}(f_t^{loc})$ reflects the correct location of sound sources thus helping the final salient object detection.

Ablation Experiments

We conduct extensive ablation studies on the ASOD60K dataset to verify the effectiveness of different components.

Backbone	ASOD60K-Test All			
	$F_\beta \uparrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$\mathcal{M} \downarrow$
Backbone	.396	.660	.714	.037
+Transformer	**.404**	**.678**	.732	**.026**
+Transformer+SPE	.403	.676	**.742**	**.026**

Table 3: Impact of visual feature extraction methods. Components are added step by step.

3D Localization	ASOD60K-Test All			
	$F_\beta \uparrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$\mathcal{M} \downarrow$
✗	.391	.667	.723	.030
✓	**.404**	**.678**	**.732**	**.026**

Table 5: Impact of 3D sound source localization branch.

Sound Type	ASOD60K-Test All			
	$F_\beta \uparrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$\mathcal{M} \downarrow$
None	.396	.660	.714	.037
Mono	.397	.662	.717	.029
Ambisonic	**.404**	**.678**	**.732**	**.026**

Table 7: Impact of sound type.

Multimodal fusion method. To investigate the effectiveness of our proposed ACF block, we conduct experiments with different multimodal fusion schemes. As shown in Table 2, we compare the ACF block equipped with multimodal attention with two baseline settings. 'None' and 'ACF (Concat)' means no multimodal fusion and simply concatenating visual and acoustic features in the ACF block respectively. In general, the model without multimodal fusion leads to an inferior performance which indicates the acoustic modality is essential to salient object detection. We notice that ACF block equipped with multimodal attention outperforms the ACF block with simply multimodal feature concatenation. Additional spherical positional encoding (SPE) brings another 0.07 F_β, 0.08 S_α and 0.1 E_ϕ gain compared to multimodal attention.

Visual feature extraction. We conduct experiments to ablate the influence of different visual feature extraction methods. We first build a baseline model that leverages ResNet-50 (He et al. 2016) backbone to extract visual features which leads to 0.396 F_β, 0.660 S_α, 0.714 E_ϕ and 0.37 $\mathcal{M}$. By employing a transformer encoder on top of the backbone, we observe non-trivial gains on all metrics. We consider the improvement comes from the strong global understanding capability of transformer. By replacing the standard positional encoding in the transformer with our spherical positional encoder, the E_ϕ metric improves 0.1 while F_β and S_α slightly drop. We consider the performance drop mainly because the pretraining is conducted on the 2D dataset.

Distillation loss weight. To investigate the influence of distillation loss weight, we conduct experiments by ablating

$\lambda_{distill}$	ASOD60K-Test All			
	$F_\beta \uparrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$\mathcal{M} \downarrow$
0.0	.389	.667	.716	.027
1.0	.402	.676	.725	.029
2.0	**.410**	.670	.715	.035
5.0	.404	**.678**	**.732**	**.026**

Table 4: Impact of distillation loss weight.

Window Size	ASOD60K-Test All			
	$F_\beta \uparrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$\mathcal{M} \downarrow$
1	.392	.670	.725	.026
2	.396	.672	.719	.027
3	**.404**	.678	**.732**	**.026**
4	.402	**.680**	.728	.028

Table 6: Impact of input window size.

different $\lambda_{distill}$. As shown in Table 4, we notice that a weight of 5 leads to the best result of 0.404 F_β, 0.678 S_α, 0.732 E_ϕ and 0.26 $\mathcal{M}$. The $\lambda_{distill} = 0$ means that no teacher branch is adopted which leads to the worst result.

3D sound source localization. To demonstrate the effectiveness of employing 3D sound source localization in VSOD, we conduct an experiment to disable the sound source localization branch in our method. The result in Table 5 indicates that 3D sound source localization from ambisonic audio can help salient object detection in panoramic scenarios.

Window size. Since temporal information is essential for the video salient object detection, we conduct experiments on different input window sizes as shown in Table 6. We notice that the window size of 3 achieves the best performance in terms of F_β, E_ϕ and $\mathcal{M}$ and the window size of 4 achieves the best result in terms of S_α.

Sound type. We conduct experiments to show the benefit of utilizing ambisonic audios. Table 7 shows that mono audio only has a trivial improvement compared to the baseline setting (no multimodal fusion). We consider this is because mono audio cannot explicitly model spatial information of the sound source.

Conclusion

In this paper, we propose a framework for audio-visual video salient object detection in panoramic scenarios. In particular, we propose an audio-visual context fusion block to enhance visual features by ambisonic audios. To better utilize the spatial information encoded in the audios, a label-guided distillation scheme is introduced to help the multimodal interaction. In addition, due to the severe distortions in the ER frame, we leverage position-agnostic attention mechanism equipped with spherical positional encoding to map each pixel back to 3D space thus capturing the true spatial location of each pixel. Notably, our method achieves the best result on the ASOD60K benchmark. Moreover, extensive study shows that ambisonic audio can help the salient object detection in panoramic videos.

References

Achanta, R.; Hemami, S.; Estrada, F.; and Susstrunk, S. 2009. Frequency-tuned salient region detection. In *2009 IEEE conference on computer vision and pattern recognition*, 1597–1604. IEEE.

Adavanne, S.; Politis, A.; Nikunen, J.; and Virtanen, T. 2018. Sound event localization and detection of overlapping sources using convolutional recurrent neural networks. *IEEE Journal of Selected Topics in Signal Processing*, 13(1): 34–48.

Bertasius, G.; Soo Park, H.; Yu, S. X.; and Shi, J. 2017. Unsupervised learning of important objects from first-person videos. In *Proceedings of the IEEE International Conference on Computer Vision*, 1956–1964.

Borji, A.; Cheng, M.-M.; Jiang, H.; and Li, J. 2015. Salient object detection: A benchmark. *IEEE transactions on image processing*, 24(12): 5706–5722.

Cai, Y.; Li, X.; Wang, Y.; and Wang, R. 2022. An Overview of Panoramic Video Projection Schemes in the IEEE 1857.9 Standard for Immersive Visual Content Coding. *IEEE Transactions on Circuits and Systems for Video Technology*.

Chao, F.-Y.; Ozcinar, C.; Wang, C.; Zerman, E.; Zhang, L.; Hamidouche, W.; Deforges, O.; and Smolic, A. 2020. Audio-visual perception of omnidirectional video for virtual reality applications. In *2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, 1–6. IEEE.

Chen, Y.-W.; Jin, X.; Shen, X.; and Yang, M.-H. 2022. Video Salient Object Detection via Contrastive Features and Attention Modules. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1320–1329.

Cheng, H.-T.; Chao, C.-H.; Dong, J.-D.; Wen, H.-K.; Liu, T.-L.; and Sun, M. 2018. Cube padding for weakly-supervised saliency prediction in 360 videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1420–1429.

Cheng, S.; Song, L.; Tang, J.; and Guo, S. 2021. Audio-Visual Salient Object Detection. In *International Conference on Intelligent Computing*, 510–521. Springer.

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. *arXiv preprint arXiv:2010.11929*.

Fan, D.-P.; Cheng, M.-M.; Liu, Y.; Li, T.; and Borji, A. 2017. Structure-measure: A new way to evaluate foreground maps. In *Proceedings of the IEEE international conference on computer vision*, 4548–4557.

Fan, D.-P.; Gong, C.; Cao, Y.; Ren, B.; Cheng, M.-M.; and Borji, A. 2018. Enhanced-alignment measure for binary foreground map evaluation. *arXiv preprint arXiv:1805.10421*.

Fan, D.-P.; Wang, W.; Cheng, M.-M.; and Shen, J. 2019. Shifting more attention to video salient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8554–8564.

Fragkiadaki, K.; Zhang, G.; and Shi, J. 2012. Video segmentation by tracing discontinuities in a trajectory embedding. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 1846–1853. IEEE.

Gao, S.-H.; Tan, Y.-Q.; Cheng, M.-M.; Lu, C.; Chen, Y.; and Yan, S. 2020. Highly efficient salient object detection with 100k parameters. In *European Conference on Computer Vision*, 702–721. Springer.

Gu, Y.; Wang, L.; Wang, Z.; Liu, Y.; Cheng, M.-M.; and Lu, S.-P. 2020. Pyramid constrained self-attention network for fast video salient object detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 10869–10876.

Guizzo, E.; Gramaccioni, R. F.; Jamili, S.; Marinoni, C.; Massaro, E.; Medaglia, C.; Nachira, G.; Nucciarelli, L.; Paglialunga, L.; Pennese, M.; et al. 2021. L3DAS21 Challenge: Machine learning for 3D audio signal processing. In *2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP)*, 1–6. IEEE.

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.

Huang, Y.; Li, X.; Wang, W.; Jiang, T.; and Zhang, Q. 2021a. Forgery Attack Detection in Surveillance Video Streams Using Wi-Fi Channel State Information. *IEEE Transactions on Wireless Communications*.

Huang, Y.; Li, X.; Wang, W.; Jiang, T.; and Zhang, Q. 2021b. Towards cross-modal forgery detection and localization on live surveillance videos. In *IEEE INFOCOM 2021-IEEE Conference on Computer Communications*, 1–10. IEEE.

Kim, C.; and Hwang, J.-N. 2002. Fast and automatic video object segmentation and tracking for content-based applications. *IEEE transactions on circuits and systems for video technology*, 12(2): 122–129.

Le, T.-N.; and Sugimoto, A. 2017. Deeply Supervised 3D Recurrent FCN for Salient Object Detection in Videos. In *BMVC*, volume 1, 3.

Li, G.; Xie, Y.; Wei, T.; Wang, K.; and Lin, L. 2018. Flow guided recurrent neural encoder for video salient object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3243–3252.

Li, H.; Chen, G.; Li, G.; and Yu, Y. 2019. Motion guided attention for video salient object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 7274–7283.

Li, X.; Wang, J.; Li, X.; and Lu, Y. 2022a. Hybrid instance-aware temporal fusion for online video instance segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 1429–1437.

Li, X.; Wang, J.; Li, X.; and Lu, Y. 2022b. Video instance segmentation by instance flow assembly. *IEEE Transactions on Multimedia*.

Li, X.; Wang, J.; Xu, X.; Li, X.; Lu, Y.; and Raj, B. 2022c. Rˆ 2VOS: Robust Referring Video Object Segmentation via Relational Multimodal Cycle Consistency. *arXiv preprint arXiv:2207.01203*.

Li, X.; Wang, J.; Xu, X.; Raj, B.; and Lu, Y. 2022d. Online Video Instance Segmentation via Robust Context Fusion. *arXiv preprint arXiv:2207.05580*.

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2117–2125.

Liu, Z.; Tan, Y.; He, Q.; and Xiao, Y. 2021. SwinNet: Swin Transformer drives edge-aware RGB-D and RGB-T salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*.

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

Lu, X.; Wang, W.; Ma, C.; Shen, J.; Shao, L.; and Porikli, F. 2019. See more, know more: Unsupervised video object segmentation with co-attention siamese networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3623–3632.

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*.

Milletari, F.; Navab, N.; and Ahmadi, S.-A. 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, 565–571. IEEE.

Nguyen, A.; Yan, Z.; and Nahrstedt, K. 2018. Your attention is unique: Detecting 360-degree video saliency in head-mounted display for head movement prediction. In *Proceedings of the 26th ACM international conference on Multimedia*, 1190–1198.

Pang, Y.; Zhao, X.; Zhang, L.; and Lu, H. 2020. Multi-scale interactive network for salient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9413–9422.

Ren, S.; Liu, W.; Liu, Y.; Chen, H.; Han, G.; and He, S. 2021a. Reciprocal transformations for unsupervised video object segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 15455–15464.

Ren, S.; Wen, Q.; Zhao, N.; Han, G.; and He, S. 2021b. Unifying Global-Local Representations in Salient Object Detection with Transformer. *arXiv preprint arXiv:2108.02759*.

Su, Y.; Deng, J.; Sun, R.; Lin, G.; and Wu, Q. 2022. A Unified Transformer Framework for Group-based Segmentation: Co-Segmentation, Co-Saliency Detection and Video Salient Object Detection. *arXiv preprint arXiv:2203.04708*.

Suzuki, T.; and Yamanaka, T. 2018. Saliency map estimation for omni-directional image considering prior distributions. In *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2079–2084. IEEE.

Tsiami, A.; Koutras, P.; and Maragos, P. 2020. Stavis: Spatio-temporal audiovisual saliency network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4766–4776.

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, H.; Jiang, X.; Ren, H.; Hu, Y.; and Bai, S. 2021. Swiftnet: Real-time video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1296–1305.

Wang, L.; Lu, H.; Wang, Y.; Feng, M.; Wang, D.; Yin, B.; and Ruan, X. 2017. Learning to detect salient objects with image-level supervision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 136–145.

Wang, W.; Shen, J.; Li, X.; and Porikli, F. 2015. Robust video object cosegmentation. *IEEE Transactions on Image Processing*, 24(10): 3137–3148.

Wang, W.; Shen, J.; and Porikli, F. 2015. Saliency-aware geodesic video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3395–3402.

Wang, W.; Shen, J.; and Shao, L. 2017. Video salient object detection via fully convolutional networks. *IEEE Transactions on Image Processing*, 27(1): 38–49.

Wang, W.; Song, H.; Zhao, S.; Shen, J.; Zhao, S.; Hoi, S. C.; and Ling, H. 2019. Learning unsupervised video object segmentation through visual attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3064–3074.

Wei, J.; Wang, S.; and Huang, Q. 2020. F^3Net: fusion, feedback and focus for salient object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 12321–12328.

Wei, J.; Wang, S.; Wu, Z.; Su, C.; Huang, Q.; and Tian, Q. 2020. Label decoupling framework for salient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13025–13034.

Wu, J.; Jiang, Y.; Sun, P.; Yuan, Z.; and Luo, P. 2022. Language as Queries for Referring Video Object Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4974–4984.

Wu, Z.; Su, L.; and Huang, Q. 2019a. Cascaded partial decoder for fast and accurate salient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3907–3916.

Wu, Z.; Su, L.; and Huang, Q. 2019b. Stacked cross refinement network for edge-aware salient object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 7264–7273.

Yan, P.; Li, G.; Xie, Y.; Li, Z.; Wang, C.; Chen, T.; and Lin, L. 2019. Semi-supervised video salient object detection using pseudo-labels. In *Proceedings of the IEEE/CVF international conference on computer vision*, 7284–7293.

Yang, C.; Zhang, L.; Lu, H.; Ruan, X.; and Yang, M.-H. 2013. Saliency detection via graph-based manifold ranking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3166–3173.

Ye, L.; Rochan, M.; Liu, Z.; and Wang, Y. 2019. Cross-modal self-attention network for referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10502–10511.

Zhang, P.; Kang, Z.; Yang, T.; Zhang, X.; Zheng, N.; and Sun, J. 2022. LGD: Label-Guided Self-Distillation for Object Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 3309–3317.

Zhang, Y.; Chao, F.-Y.; and Zhang, L. 2021. ASOD60K: An Audio-Induced Salient Object Detection Dataset for Panoramic Videos. *arXiv preprint arXiv:2107.11629*.

Zhang, Z.; Xu, Y.; Yu, J.; and Gao, S. 2018. Saliency detection in 360 videos. In *Proceedings of the European conference on computer vision (ECCV)*, 488–503.

Zhang, Z.; Yu, C.; and Crandall, D. 2019. A self validation network for object-level human attention estimation. *Advances in Neural Information Processing Systems*, 32.

Zhao, C.; Li, X.; Dong, S.; and Younes, R. 2022a. Self-supervised Multi-Modal Video Forgery Attack Detection. *arXiv preprint arXiv:2209.06345*.

Zhao, X.; Pang, Y.; Zhang, L.; Lu, H.; and Zhang, L. 2020. Suppress and balance: A simple gated network for salient object detection. In *European conference on computer vision*, 35–51. Springer.

Zhao, Y.; Guo, X.; and Lu, Y. 2022. Semantic-aligned Fusion Transformer for One-shot Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7601–7611.

Zhao, Y.; Li, Z.; Guo, X.; and Lu, Y. 2022b. Alignment-guided Temporal Attention for Video Action Recognition. *arXiv preprint arXiv:2210.00132*.

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*, 386–403. Springer.

Zhou, T.; Li, J.; Wang, S.; Tao, R.; and Shen, J. 2020. Matnet: Motion-attentive transition network for zero-shot video object segmentation. *IEEE Transactions on Image Processing*, 29: 8326–8338.

Zhu, Y.; Zhai, G.; and Min, X. 2018. The prediction of head and eye movement for 360 degree images. *Signal Processing: Image Communication*, 69: 15–25.