Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Audio-visual saliency prediction with multisensory perception and integration

Jiawei Xie [a], Zhi Liu [a,b,*], Gongyang Li [a,b], Yingjie Song [a]

[a] Shanghai Institute for Advanced Communication and Data Science, School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China

[b] Wenzhou Institute of Shanghai University, Wenzhou 325000, China

ARTICLE INFO

Keywords:
Audio-visual saliency prediction
Audio-visual fusion
Image saliency prediction
Self-supervised learning

ABSTRACT

Audio-visual saliency prediction (AVSP) is a task that aims to model human attention patterns in the perception of auditory and visual scenes. Given the challenges associated with perceiving and combining multi-modal saliency features from videos, this paper presents a multi-sensory framework for AVSP. This framework is designed to extract audio, motion and image saliency features and integrate them effectively, which can then serve as a general architecture for the AVSP task. To obtain multi-sensory information, we develop a three-stream encoder that extracts audio, motion and image saliency features. In particular, we utilize a pre-trained encoder with knowledge related to image saliency to extract saliency features for each frame. The image saliency features are then incorporated with motion features using a spatial attention module. For motion features, 3D convolutional neural networks (CNNs) like S3D are commonly used in AVSP models. However, these networks are unable to effectively capture the global motion relationship in videos. To tackle this problem, we incorporate Transformer- and MLP-based motion encoders into the AVSP models. To learn joint audio-visual representations, an audio-visual fusion block is exploited to enhance the correlation between audio and visual motion features under the supervision of a cosine similarity loss in a self-supervised manner. Finally, a multi-stage decoder integrates audio, motion and image saliency features to generate the final saliency map. We evaluate our methods on six audio-visual eye-tracking datasets. Experimental results demonstrate that our method achieves compelling performance compared to the state-of-the-art methods. The source code is available at https://github.com/oraclefina/MSPI.

1. Introduction

Nowadays, humans are frequently exposed to multi-sensory and cross-modal stimuli from videos on televisions and mobile phones. While watching videos, humans perceive auditory and visual information simultaneously and can quickly pay attention to places of interest. The understanding of this human visual attention mechanism is an active research field in computer vision. Recently, with the advent of deep learning, image saliency prediction (ISP) models [5,18,22,37,44,54,55,70,81,89,92,98,100] and video saliency prediction (VSP) models [47,61,73,96,97,105,107] have made a significant progress in predicting visual attention with static images and spatiotemporal visual features. However, the audio-visual attention modeling based on deep learning methods is still in early stages, because most video saliency models primarily focus on visual information, and the collection of large-scale eye-tracking datasets for videos [38,60,95] is typically done in a soundless environment.

In videos, visual streams are naturally accompanied by auditory streams, which have been embodied in many video-related works [1,4,106]. Unlike the fact that the human brain can perceive and process these multi-sensory stimuli in a sophisticated and elegant way, most of audio-visual models rely on two-stream neural networks to process video frames and audio for learning audio-visual features. As shown in Fig. 1 (a), current AVSP models [36,67,83,87,108] also utilize two-stream networks to simultaneously acquire dynamic visual feature representations and auditory semantic feature representations. For the audio stream, audio encoders [6,12,33] for large-scale audio classification are employed to extract audio feature representation. For the visual stream, 3D CNN-based motion encoders obtain visual dynamic features over successive video frames. However, on the one hand, 3D

[*] Corresponding author.
E-mail address: liuzhisjtu@163.com (Z. Liu).

https://doi.org/10.1016/j.imavis.2024.104955
Received 27 October 2023; Received in revised form 21 December 2023; Accepted 17 February 2024
Available online 23 February 2024

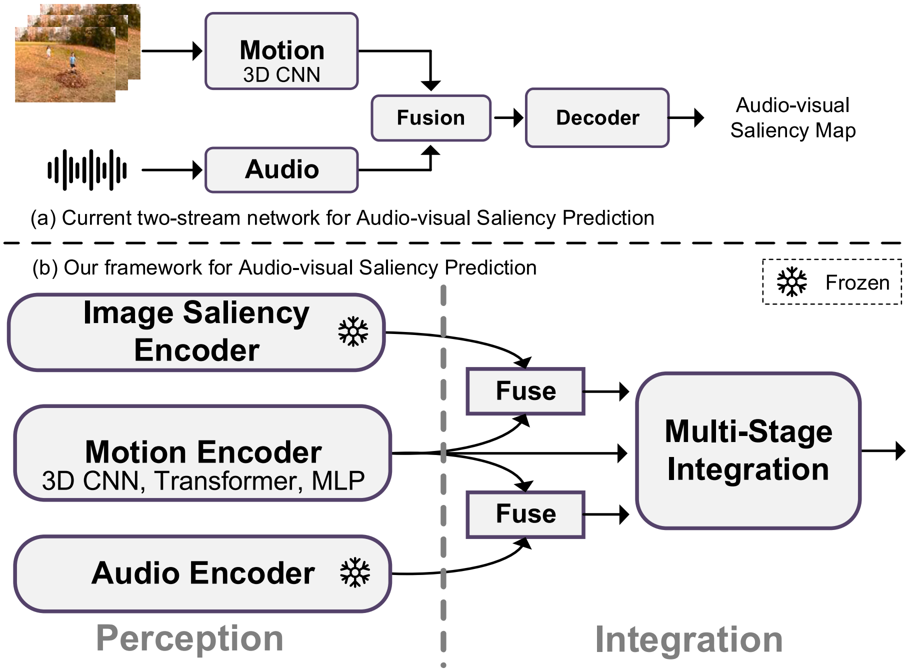

Fig. 1. Architectures of the audio-visual saliency prediction models. (a) The current two-stream network for audio-visual saliency prediction. (b) The overview of our multi-sensory framework for audio-visual saliency prediction.

convolutions cannot effectively capture the long-range motion relationships among frames. On the other hand, these AVSP models often overlook the frame-level static saliency features as temporal dimensions gradually decrease by temporal pooling operations. Based on the huge success in ISP models, Min et al. [63] directly integrated audio-visual saliency maps with static saliency maps produced by ISP models to promote the performance under audio-visual scenes. But the potential of the ISP models has not been fully explored for AVSP models.

How to integrate auditory and visual features and acquire audio-visual correspondence representation for AVSP models is the key issue. Some works [36,67,87] attempt to locate the sounding object through different audio-visual fusion methods. They utilize pooling operations on audio and visual features before audio-visual fusion. These pooling operations either removes the temporal information or spatiotemporal information of these features. To explore explicit audio-visual correspondence cues, Chen et al. [11] manually annotated the audio-visual consistency labels of videos and incorporated an audio-visual consistency classifier to regulate the integration of audio and visual features. However, this approach requires significant efforts to label video frames. Recently, self-supervised learning has gained popularity and promoted the advancement of audio-visual learning [2,16,29,68]. For the AVSP task, Xiong et al. [102] introduced a consistency-aware predictive coding module to iteratively minimize the distance between audio and visual feature embeddings, which shows a promising future for introducing self-supervised methods into AVSP models.

Based on the above observations, we propose a multi-sensory framework for audio-visual saliency prediction, which aims to perceive motion, auditory and image saliency information, and then combine these modalities based on their respective characteristics. The overview of our framework is depicted in Fig. 1 (b), which is divided into perception and integration stages. To simulate multi-sensory perception, we construct a three-stream network that incorporates motion, audio and image saliency encoders to extract multi-modal features. In particular, to fully realize the potential of ISP models, we obtain frame-level saliency features from an image saliency encoder pretrained on image saliency datasets [40,41], pursuing the later feature-level fusion in latent space. To address the issue of that the 3D CNN-based motion encoders currently used in AVSP models cannot well capture long-range motion saliency features, we introduce Transformer- and MLP-based motion encoders into AVSP. Transformer structure can capture long-range relationships by the self-attention operation and MLP structure can obtain global information through fully-connected layers. For the integration stage, multi-modal fusion methods are based on the characteristics of each modality. The extracted image saliency features are incorporated into motion features through a spatial attention module. To learn audio-visual correspondence representations in a self-supervised manner, we build a audio-visual fusion block that comprises Transformer encoder layers, projection layers and predictor layers. We employ a cosine similarity loss to align the audio and visual features based on the *stopgrad* operation from [15]. Finally, a multi-stage decoder is designed to fully integrate audio and visual features. We conducted experiments on six widely-used audio-visual eye-tracking datasets and evaluated our multi-sensory framework using 3D CNN [26,27,101], Transformer [52,53,57] and MLP [104] as motion encoders. Experimental results show that our model achieves compelling performance compared to the state-of-the-art methods.

In summary, the main contributions of this work are detailed as follows:

- We propose a multi-sensory framework for audio-visual saliency prediction that can perceive motion, audio and image saliency features and integrate them effectively based on their characteristics. Furthermore, to demonstrate the general use of the proposed framework and the importance of the long-range modeling, we evaluate our framework using 3D CNN-, Transformer- and MLP-based motion encoders for AVSP. Experiments show that the long-range motion capture structures are beneficial for AVSP.
- We introduce an image saliency encoder into the architecture of AVSP. We simplify the structure of ISP models and analyze the inherent operating behavior of image saliency features in our framework. With prior knowledge of image visual saliency, our model gains the ability to extract frame-level saliency features and adjust the spatio-temporal saliency features through a spatial attention module.
- We propose an audio-visual fusion block to learn audio-visual correspondence relationships in a self-supervised manner. Transformer

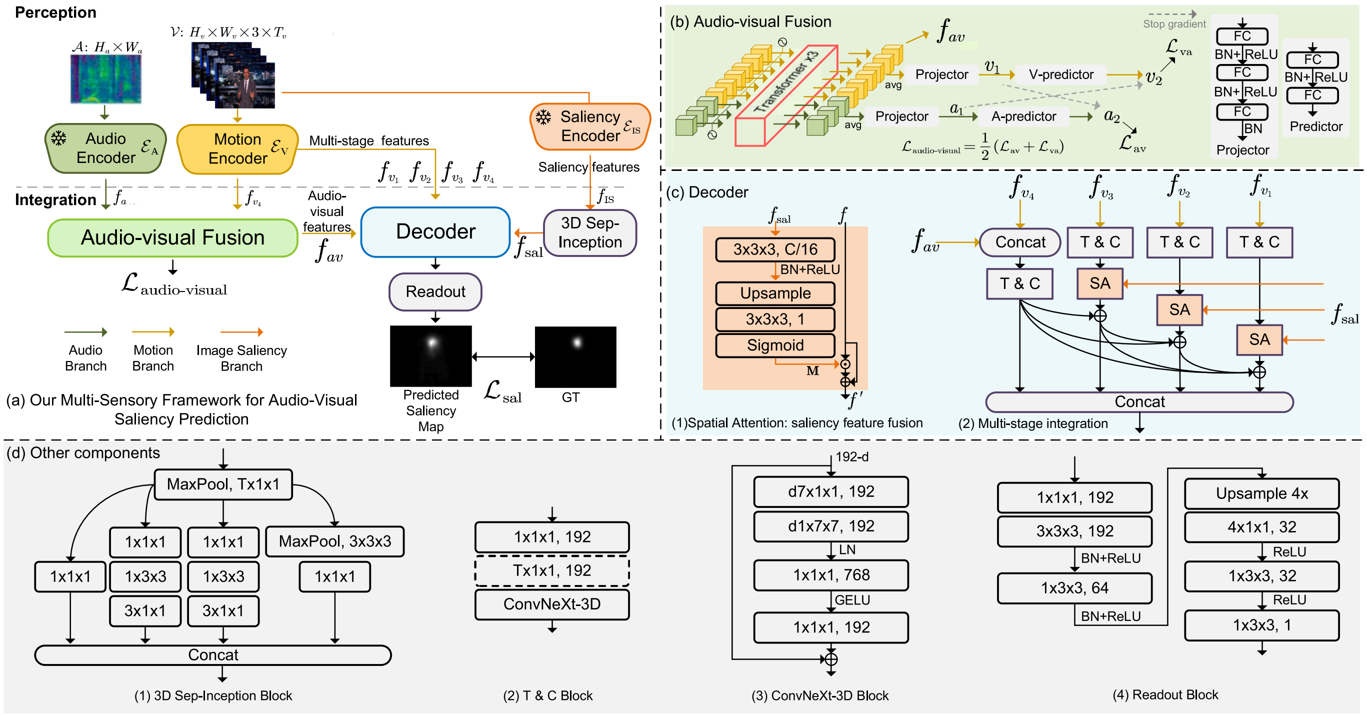

Fig. 2. The architecture of our multi-sensory framework for audio-visual saliency prediction. (a) The overview of our method, which consists of a three-stream encoder, an audio-visual fusion block, a decoder for further feature fusion and a Readout block to generate saliency maps. (b) The detailed audio-visual fusion block. (c) The detailed decoder and the SA block. (d) The structure of other four components, containing a 3D Sep-Inception Block [101], T&C Block, 3D version of ConvNeXt Block [56] and Readout Block.

layers and a symmetrized cosine similarity loss are used for the representation learning between audio and visual features.

2. Related work

In this section, we briefly review the literature on visual saliency prediction, audio-visual saliency prediction and self-supervised audio-visual learning.

2.1. Visual saliency prediction

Compared to detecting salient objects [48–51,71,93,94], visual saliency prediction aims to predict human fixations when human observe visual stimulus like images and videos. According to the type of input, visual saliency works can be categorized into image saliency prediction and video saliency prediction. For image saliency prediction, traditional models depend on hand-crafted features including low-level features (e.g., color, contrast and orientation) [7,25,35] and high-level features (e.g., text and faces) [8,85]. Recently, with the prevalence of deep learning, CNN-based approaches have been introduced into saliency prediction and have achieved great progress. Many image saliency models are built based on an encoder-decoder structures, where the encoders are typically pre-trained CNN backbones while the decoder strategies are various. Some works [24,45,76,81,92] aggregate hierarchical intermediate maps from different encoder layers to generate saliency maps. Some works [18,69,103] build independent encoders and decoders with no feature sharing between them.

As for video saliency prediction, the majority of the models employ long-term temporal modeling structures, such as LSTM [13,39,99], RNN [23] and ConvGRU [47], to generate saliency maps over successive frames. Some works [36,61,97] build models based on 3D convolutions to learn spatio-temporal saliency features. Recently, Ma et al. [59] and Zhou et al. [107] proposed transformer-based models to learn long-term spatiotemporal features for video saliency prediction, achieving outstanding performance. In this paper, we leverage the knowledge pertaining to image saliency from ISP models to acquire frame-level saliency features. In order to effectively extract long-range motion features, we evaluate our multi-sensory framework on various motion encoders, including 3D CNN-, Transformer- and MLP-based video backbones.

2.2. Audio-visual saliency prediction

Psychological studies [72,90] have proven that the auditory modality can influence the perception of the visual modality. Song et al. [80] analyzed eye movement of humans under different types of sound, revealing that humans in the audio-visual condition exhibited more frequent eye movements. Coutrot et al. [19,20] studied human visual behavior in social multi-modal scenarios. Recent works [10,63,83,91] have started to employ deep learning-based techniques to model the audio-visual attention mechanism and propose approaches to incorporate visual and audio modalities for AVSP. Tsimami et al. [87] introduced a bilinear fusion operation to integrate multi-modality features. Zhu et al. [108] utilized a canonical correlation analysis method to capture the correspondence between multi-modal information streams. Chen et al. [11] manually annotated the audio-visual consistency labels for videos and designed a classifier to control the audio-visual integration, which is able to learn explicit audio-visual correspondence relationships. Xiong et al. [102] introduced a consistency-aware predictive coding module to iteratively minimize the distance between audio and visual feature embeddings in a self-supervised manner, which shows a promising future for incorporating self-supervised methods into AVSP models.

2.3. Self-supervised audio-visual learning

Recently, self-supervised learning methods like contrastive learning [14,15,30] have been introduced into audio and visual representation

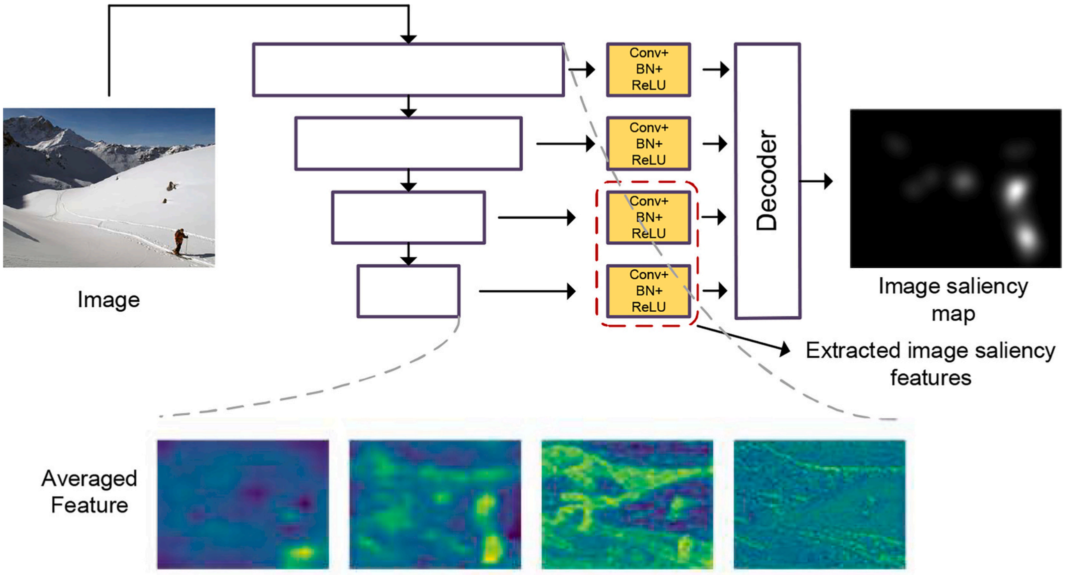

Fig. 3. The architecture of the ISP model and the visualization of averaged feature in each block from the encoder. The red dotted box here represents the image saliency features that are used for the later fusion in our framework. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

learning [3,28,34,66]. Most approaches utilize single-modality backbones to generate representations of each modality, which are then optimized with self-supervised loss [75,78]. Recent works attempt to learn joint audio-visual representations using a single backbone [65,74]. Shvetsova et al. [79] proposed a multi-modal, modality-agnostic fusion transformer to exchange information from audio, video and text and trained the model with a combinatorial loss to obtain multi-modal shared embedding space. Gong et al. [29] combined contrastive learning and masked data modeling to learn a joint audio-visual representation, which can capture both modality-unique and audio-visual paired information. These studies show that the contrastive learning methods can obtain the multi-modal correlation feature representation. Thus, in this paper, we propose to utilize a symmetrized cosine similarity loss to supervise audio-visual feature representation learning with transformer encoder, projector and predictor layers for AVSP.

3. Approach

In this section, we first introduce the overview of our multi-sensory framework for audio-visual saliency prediction, and then describe the detailed structure of each module and loss function.

3.1. Overview

Our framework contains two stages, i.e., a perception stage with a three-stream encoder for audio, motion and image saliency feature extraction, and a integration stage with multiple fusion blocks and a decoder to fuse audio-visual, motion-image saliency and multi-stage features, respectively. The detailed architecture is illustrated in Fig. 2. Given a frame sequence $\mathscr{V} \in \mathbb{R}^{H_v \times W_v \times 3 \times T_v}$ and an audio spectrogram $\mathscr{A} \in \mathbb{R}^{H_a \times W_a}$, the audio-visual saliency model generates the saliency map. First, the audio and visual streams are encoded with three feature extractors $\mathscr{E}_\mathrm{A}$, $\mathscr{E}_\mathrm{V}$ and $\mathscr{E}_\mathrm{IS}$. Second, the extracted audio and motion features are fed into an audio-visual fusion block to learn audio-visual correspondence features, while the extracted image saliency features are taken into a 3D Sep-Inception block to enhance temporal saliency features for the later fusion with motion features. Next, the decoder integrates image saliency, audio-visual and multi-scale visual features. Finally, a readout block is used to generate the final saliency map.

3.2. Three-stream encoder

To extract audio, motion and image saliency features for AVSP, a three-stream network with $\mathscr{E}_\mathrm{A}$, $\mathscr{E}_\mathrm{V}$ and $\mathscr{E}_\mathrm{IS}$ is established. As depicted in Fig. 2 (a), audio feature representation f_a is extracted through an audio encoder $\mathscr{E}_\mathrm{A}$:

$$f_a = \mathscr{E}_\mathrm{A}(\mathscr{A}) \in \mathbb{R}^{h_a \times w_a \times C_a} \tag{1}$$

Visual input is encoded in a pyramid motion features extractor $\mathscr{E}_\mathrm{V}$ and an image saliency extractor $\mathscr{E}_\mathrm{IS}$:

$$f_{v_i} = \mathscr{E}_\mathrm{V}(\mathscr{V}) \in \mathbb{R}^{h_{v_i} \times w_{v_i} \times C_{v_i} \times T_{v_i}}, i = 1, 2, 3, 4 \tag{2}$$

$$f_\mathrm{IS} = \mathscr{E}_\mathrm{IS}(\mathscr{V}) \in \mathbb{R}^{h_\mathrm{IS} \times w_\mathrm{IS} \times C_\mathrm{IS} \times T_v} \tag{3}$$

where i represents different stages of the motion extractor. For $\mathscr{E}_\mathrm{A}$, we use a ResNet-18 [32] variant that is pre-trained on VGGSound [12] for sound classification. For $\mathscr{E}_\mathrm{V}$, we conduct experiments on multiple pyramid video backbones of different structures, including three 3D CNN-based backbones (S3D [101], X3D [26] and SlowFast [27]), one MLP-based backbone (MorphMLP [104]) and three Transformer-based backbones (VideoSwin [57], Uniformer [52] and MViTv2 [53]), which are pre-trained on Kinetics 400 [42]. The Transformer- and MLP-based backbones are able to effectively capture long-term motion features. As for $\mathscr{E}_\mathrm{IS}$, we adopt the ImageNet-pretrained MobileNet-V2 [77] and ConvNeXt-T [56] as ISP models, which are then finetuned on image eye-tracking datasets, i.e., SALICON [40] and MIT1003 [41]. During the training process for AVSP, the weights of the audio and image saliency encoders are frozen.

3.3. Incorporating image saliency features

To obtain static saliency features, an image saliency model finetuned on image saliency datasets can provide comprehensive image saliency information. Image saliency prediction models [24,45,76,81] propose a variety of designs to capture, enhance and fuse saliency features, but they all have a typical encoder-decoder architecture. The detailed designs of decoders are various, but the encoders are basically the same and features in the encoders can easily be obtained. Thus, we simplify the ISP architecture as the one in Fig. 3 with Conv+BN + ReLU as a

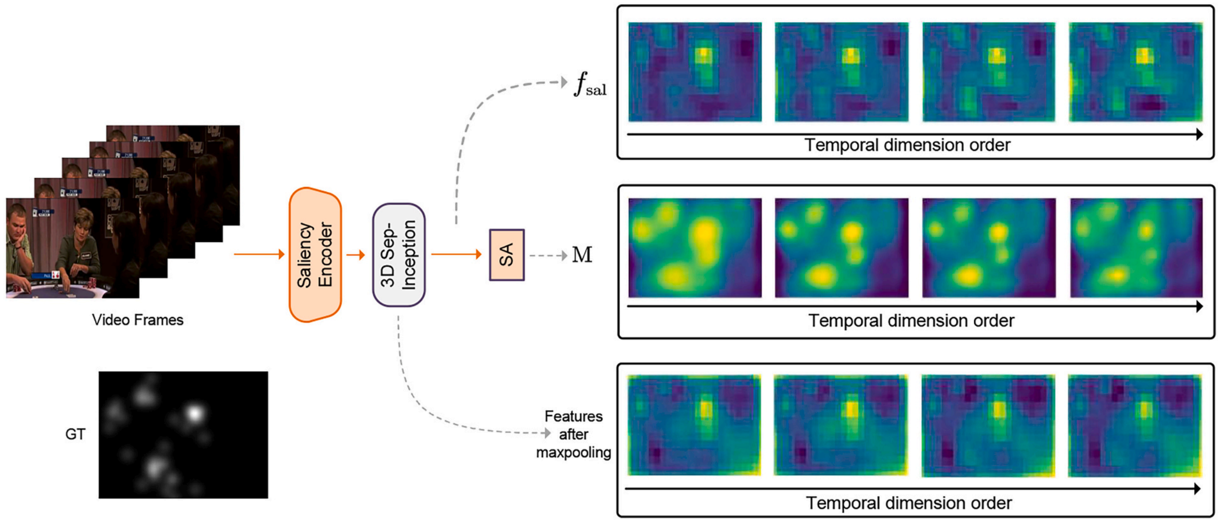

Fig. 4. Visualization of image saliency features from the 3D Sep-Inception block and SA block.

transition layer between the encoder and the decoder. Features in the deep stages of ISP models, which include semantic information such as objects, faces and text, are crucial for explaining free viewing behavior in natural scenes [46,82]. The features in the last two stages with a large receptive field contain rich semantic information. From the averaged features in Fig. 3, the features in the deep stages highlight the salient regions and are more similar to the saliency map. Therefore, we obtain the image saliency features from the last two stages of the encoder as shown in the red dotted box in Fig. 3. Since the image saliency encoder already possesses the ability to highlight the salient regions in images, it is beneficial to freeze its weights during training, which also reduces the consumption of graphic memory.

The extracted saliency features from ISP encoder contains rich saliency information, but are in weak representation of temporal properties. To better incorporate the image saliency features into multi-stage visual features from $\mathscr{E}_{\text{V}}$, a sep-inception block [101] is utilized. This block enhances the spatiotemporal relationship and obtains the temporal enhanced image saliency features $f_{\text{sal}} \in \mathbb{R}^{h_{\text{IS}} \times w_{\text{IS}} \times 512 \times T}$ ($T = 4$) for later fusion. The detailed structure is shown in Fig. 2 (d) (1). The extracted image saliency features f_{IS} are passed through a temporal MaxPool operation first to reduce the temporal dimension. Later, the enhanced image saliency features f_{sal} are passed to the decoder for the saliency feature fusion. The detailed process of feature fusion is shown in Fig. 2 (c) (1). We utilize a spatial attention (SA) module to incorporate image saliency features f_{sal} into multi-stage motion features from f_{v_1}, f_{v_2} and f_{v_3}. As illustrated in Fig. 2 (c) (1), for inputs f_{sal} and $f \in \mathbb{R}^{h \times w \times C \times T}$, the channel of features f_{sal} is first squeezed to 32 through a 3D Conv-BN-ReLU layer. We adopt a trilinear interpolation to align the spatial size between f_{sal} and f. Next, a $3 \times 3 \times 3$ convolutional layer and a sigmoid activation layer are used to generate a spatial attention mask $\mathbf{M} \in \mathbb{R}^{h \times w \times 1 \times T}$. Finally, the spatial relationship of f is refined by the attention mask and a residual connection is applied to obtain the fused features. The whole process is as follows:

$$\begin{aligned} &\mathbf{M} = \sigma(\text{Conv3D}(\text{Up}(\text{ReLU}(\text{BN}(\text{Conv3D}(f_{\text{sal}})))))), \\ &f^{'} = f + f \odot \mathbf{M} \end{aligned} \tag{4}$$

where $\sigma(\cdot)$ represents the sigmoid function, $\odot$ means the element-wise multiplication.

We visualize image saliency features in our framework in Fig. 4. Our findings are as follows: the image saliency features from the saliency encoder highlight the salient regions but have a weak expression in the temporal relationship (in the third row); enhancing the image saliency features with spatiotemporal enhancement blocks can help adjust the saliency features to better align with motion features (in the first row); our SA block can provide temporal saliency mask $\mathbf{M}$, as it gradually adjusts spatial attention weights and concentrates on salient regions (in the second row).

3.4. Audio-visual fusion

To integrate audio-visual features, we adopt three transformer encoder layers with 4 heads and 512 embedding dimension to jointly process the extracted audio and visual features. The audio input f_a is flatten to audio tokens $\mathbf{a} \in \mathbb{R}^{N_a \times C_{av}}$. The visual input f_{v_4} is flattened and through a linear layer to obtain visual tokens $\mathbf{v} \in \mathbb{R}^{N_v \times C_{av}}$. Then tokens $\mathbf{a}$, $\mathbf{v}$ are added with sinusoidal positional embedding [88] and concatenated together to obtain $\mathbf{X}_{av} \in \mathbb{R}^{N \times C_{av}}$ ($N = N_a + N_v$). Transformer layers process $\mathbf{X}_{av}$ to get the final audio-visual embedding $\mathbf{X}^{'}_{av}$. The whole process is defined as follows:

$$\begin{aligned} \mathbf{a} &= \text{Flatten}(f_a) + \mathbf{E_a}, \\ \mathbf{v} &= \text{Linear}\left(\text{Flatten}\left(f_{v_4}\right)\right) + \mathbf{E_v}, \\ \mathbf{X}_{av} &= \text{Concat}(\mathbf{a}, \mathbf{v}), \\ \mathbf{X}^{'}_{av} &= \text{Transformer}_{\times 3}(\mathbf{X}_{av}) \end{aligned} \tag{5}$$

where $\mathbf{E_a}$ and $\mathbf{E_v}$ are the positional embedding of audio and visual features, respectively.

To learn audio-visual cues in a self-supervised manner, we follow the method used in [15,78]. The method can perform contrastive learning on different views of an image or multi-modality features of audio and vision using a symmetrized cosine similarity loss to learn the feature embeddings with *stopgrad*. The primary objective of *stopgrad* is to prevent model collapsing when models are trained from scratch. Given our utilization of pre-trained encoders, the training of our models is stable. Thus, to maintain the consistency of the learning method and to learn audio-visual features, we minimize the cosine embedding distance $\mathscr{D}$ between audio modality and visual modality with *stopgrad*. Given two vectors p and z, in which p represents the output vector from a predictor head and z is the output vector from a projector head or directly from the encoder followed by the *stopgrad* operation, we obtain the following equation:

$$\mathscr{D}(p, stopgrad(z)) = -\frac{p}{\|p\|_2} \cdot \frac{z}{\|z\|_2} \tag{6}$$

where $\|\cdot\|_2$ is l_2 -norm. As shown in Fig. 2 (b), processed by the three transformer layers, the concatenated features $\mathbf{X}^{'}_{av}$ are first split and

Table 1
Quantitative performance comparison with other AVSP methods on the test sets of DIEM, ETMD, SumMe, AVAD, Coutrot1 and Coutrot2 datasets. All the methods utilize the same dataset splitting strategy as [87]. Note that AViNet and TSFP-Net are pre-trained on a large-scale video eye-tracking dataset [95] first, while other AVSP models are directly trained on the six audio-visual eye-tracking datasets. The best results are **highlighted** [9].

Model	Visual Input	DIEM [64]					ETMD [43, 86]					SumMe [31, 86]					AVAD [62]					Coutrot1 [20]					Coutrot2 [21]				
		CC↑	NSS↑	AUC↑	sAUC↑	SIM↑	CC↑	NSS↑	AUC↑	sAUC↑	SIM↑	CC↑	NSS↑	AUC↑	sAUC↑	SIM↑	CC↑	NSS↑	AUC↑	sAUC↑	SIM↑	CC↑	NSS↑	AUC↑	sAUC↑	SIM↑	CC↑	NSS↑	AUC↑	sAUC↑	SIM↑
AViNet [36]	$32 \times 224 \times 384$	.6310	2.50	.8970	**.7200**	.4970	.5660	3.05	.9280	.7370	.4040	.4710	2.42	.8990	**.6990**	.3460	.6830	3.74	.9310	**.6610**	.4940	.5560	2.68	.8870	**.6360**	.4260	.7530	5.81	.9510	**.7430**	.4860
TSFP-Net [9]	$32 \times 192 \times 352$	.6510	2.62	.9060	-	.5270	.5760	3.07	.9320	-	.4280	.4640	2.30	.8940	-	.3600	**.7040**	3.77	.9320	-	.5210	.5710	2.73	.8950	-	.4470	.7430	5.31	.9590	-	.5280
STAViS [87]	16×112^2	.5795	2.26	.8838	.6741	.4824	.5690	2.94	.9316	.7317	.4251	.4220	2.04	.8883	.6562	.3373	.6086	3.18	.9196	.5936	.4578	.4722	2.11	.8686	.5847	.3935	.7349	5.28	.9581	.7106	.5111
STAViS* [87]	$16 \times 224 \times 384$	.5798	2.29	.8853	.6696	.4891	.5849	3.09	.9343	.7385	.4386	.4267	2.10	.8883	.6561	.3429	.6539	3.53	.9246	.6071	.4905	.4965	2.29	.8677	.5804	.4136	.7316	5.64	.9611	.7114	.5373
Ning et al. [67]	16×112^2	.5924	2.33	.8941	.6982	.4917	.5664	3.05	.9351	.7406	.4325	.4392	2.25	.8945	.6712	.3428	.6262	3.57	.9251	.6203	.4820	.4985	2.44	.8798	.6042	.4154	.7481	5.45	.9537	.7294	.5266
CASP-Net [102]	$16 \times 224 \times 384$	.6490	2.58	.9040	-	.5360	.6160	3.31	.9390	-	**.4760**	.4860	2.52	.9040	-	.3770	.6850	3.77	.9320	-	.5280	.5600	2.66	.8870	-	.4530	.7660	6.11	.9630	-	.5730
Ours (S3D [101])	16×224^2	.6495	2.59	.9057	.6942	.5265	.5919	3.15	.9353	.7457	.4415	.4761	2.45	.8992	.6822	.3672	.6875	3.82	.9318	.6118	.5218	.5403	2.58	.8895	.6063	.4324	.7658	6.21	.9612	.7151	.5528
Ours (S3D [101])	$16 \times 224 \times 384$	.6532	2.62	.9070	.6958	.5303	.6014	3.24	.9365	.7499	.4544	.4817	2.49	.9011	.6824	.3721	.6969	3.87	.9350	.6124	.5290	.5665	2.76	.8950	.6129	.4525	.7827	6.28	.9628	.7214	.5726
Ours (MViTv2-S [53])	16×224^2	.6592	2.65	.9090	.6991	.5372	.6147	3.27	**.9405**	.7525	.4633	.4958	2.55	.9075	.6942	.3758	.6927	3.86	.9356	.6120	.5247	.5832	2.82	.8946	.6162	.4594	.7691	6.23	.9622	.7115	.5686
Ours (MViTv2-S [53])	$16 \times 224 \times 384$	**.6711**	**2.71**	**.9105**	.7053	**.5455**	**.6238**	**3.35**	.9401	**.7570**	.4713	**.5028**	**2.61**	**.9091**	.6977	**.3804**	.7023	**3.92**	**.9364**	.6149	**.5334**	**.6057**	**2.98**	**.8996**	.6252	**.4731**	**.7998**	**6.40**	**.9648**	.7258	**.5961**

averaged into $a \in \mathbb{R}^{C_{av}}$ and $v \in \mathbb{R}^{C_{av}}$, and audio-visual features f_{av} are used for later multi-stage integration. Then a and v are fed to projectors which are MLP layers with a batchnorm layer as the output to generate feature representations a_1 and v_1. Next, A-predictor and V-predictor are utilized to obtain the audio vector a_2 and visual vector v_2. Finally, we use the negative cosine loss to minimize the cross-modal distance among these features. Thus, a symmetrized cosine similarity loss is defined as follows:

$$\mathscr{L}_{\text{audio-visual}} = \frac{1}{2}\left(\mathscr{D}(a_2, stopgrad(v_1)) + \mathscr{D}(v_2, stopgrad(a_1))\right) \tag{7}$$

3.5. Decoder

Multi-stage integration can effectively improve the performance for saliency prediction [17,36]. Inspired by this, our decoder is based on a multi-stage fusion structure. As presented in Fig. 2 (c), the visual inputs f_{v_1}, f_{v_2} and f_{v_3} are first processed by a T&C block (Temporal and Channel lateral block) to reduce the channel dimension and align the temporal dimension, while the inputs f_{v_4} and f_{av} are first concatenated and then passed into the T&C block. The detailed structure of T&C block is shown in Fig. 2 (d) (2). $1 \times 1 \times 1$ and T-convolutional layers are used for channel and temporal dimension adjustment. The dotted line represents that this operation is not performed if the temporal dimension of input satisfies the temporal setting (Temporal dimension is 4). These two operations can ease the burden of computing, but result in a loss of spatiotemporal information. Thus, an enhancement block is exploited to enhance spatiotemporal features and we adopt a 3D version of ConvNeXt block [56] (in Fig. 2 (d) (3)), by replacing 2D convolutional layers with 3D convolutional layers and utilizing depthwise separable convolutional layers to enhance temporal and spatial features. After T&C blocks, the features from f_{v_1}, f_{v_2} and f_{v_3} are fused with image saliency features f_{sal} as discussed in Section 3.3. Then the enhanced multi-stage features are fused through addition with dense connections from deep stages to early stages. The saliency map is obtained using the Readout Block (in Fig. 2 (d) (4)) that consists of 3D Conv, BN, ReLU and Upsampling layers. Finally, the predicted saliency map is normalized by a softmax operation.

3.6. Loss function

For saliency prediction, we adopt the combination of Pearson's Correlation Coefficient (CC) and Kullback–Leibler Divergence (KL) as the saliency loss $\mathscr{L}_{\text{sal}}$, in which CC estimates the correlation between two variables and KL measures the differences between two probability distributions. The process is as follows:

$$CC(\mathbf{P}, \mathbf{Q}) = \frac{\sigma(\mathbf{P}, \mathbf{Q})}{\sigma(\mathbf{P}) \times \sigma(\mathbf{Q})}, \tag{8}$$

$$KL(\mathbf{P}, \mathbf{Q}) = \sum_i \mathbf{Q}_i log\left(\varepsilon + \frac{\mathbf{Q}_i}{\mathbf{P}_i + \varepsilon}\right), \tag{9}$$

$$\mathscr{L}_{\text{sal}}(\mathbf{P}, \mathbf{Q}) = KL(\mathbf{P}, \mathbf{Q}) - CC(\mathbf{P}, \mathbf{Q}), \tag{10}$$

where $\mathbf{P}$ and $\mathbf{Q}$ are the predicted and ground-truth saliency maps, respectively, i is the pixel index, ε is a regularization term and $\sigma(\mathbf{P}, \mathbf{Q})$ is the covariance of $\mathbf{P}$ and $\mathbf{Q}$.

For audio-visual learning, as discussed in Section 3.4, the symmetrized negative cosine similarity is exploited to supervise the audio-visual correspondence learning. Thus, we deploy the total loss function as follows:

$$\mathscr{L}_{total}(\mathbf{P}, \mathbf{Q}) = \mathscr{L}_{\text{sal}}(\mathbf{P}, \mathbf{Q}) + \lambda \cdot \mathscr{L}_{\text{audio-visual}}, \tag{11}$$

where λ is the weight of audio-visual loss and is set to 1 by default.

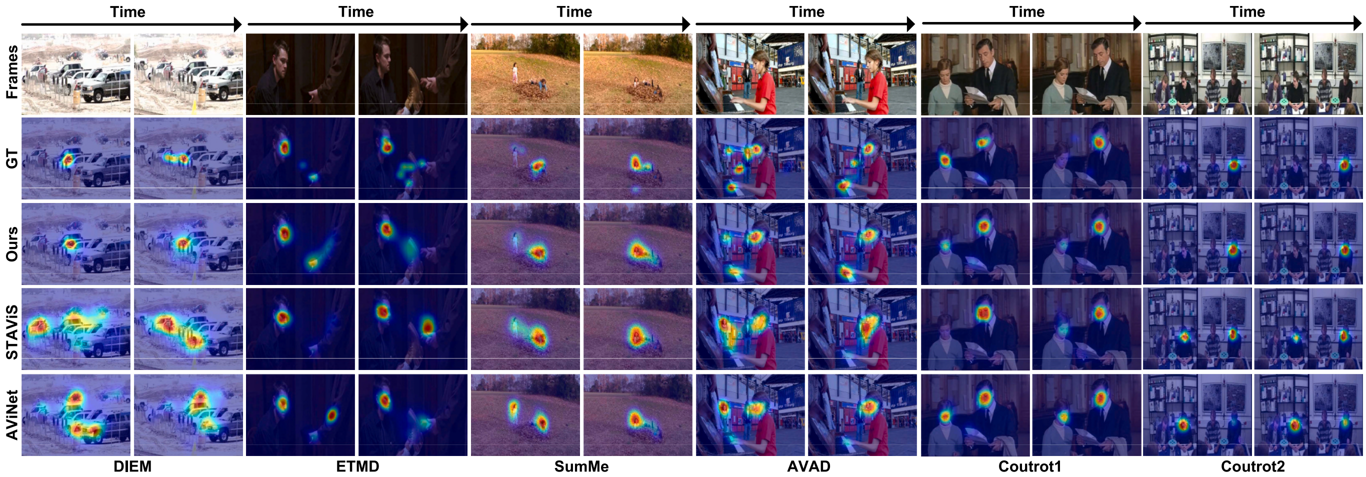

Fig. 5. Visual comparison with other state-of-the-art AVSP models.

4. Experiment

In this section, we first describe the datasets and evaluation metrics in Section 4.1 and Section 4.2. Then we introduce the details of implementation in Section 4.3. In Section 4.4 and Section 4.5, we compare our method with the state-of-the-art AVSP models and analyze the importance of image saliency features, audio-visual features and motion features to our framework by ablation studies.

4.1. Datasets

In the experiments, we use the following six audio-visual eye-tracking datasets:

DIEM [64] dataset contains 84 video clips covering music videos, game trailers, advertisement, news clips, movie trailers, commercials and documentaries.

ETMD [43,86] dataset is composed of 12 video clips from several Hollywood movies.

SumMe [31,86] dataset consists of 25 unstructured video clips which are acquired in a controlled psychological experiment.

AVAD [62] dataset contains 45 video clips with 5–10 s, which cover several audio-visual scenarios, i.e., playing basketball, playing the panio, etc.

Coutrot1 [20] contains 60 video clips with 4 visual categories: one moving object, landscapes, several moving objects and faces.

Coutrot2 [21] contains 15 video clips from a meeting with 4 persons.

4.2. Evaluation metrics

Following the former works [67,87,91,106], we utilize five metrics containing NSS (Normalized Scanpath Saliency), AUC (Area under ROC Curve), sAUC (Shuffled AUC), CC (Pearson's Correlation Coefficient) and SIM (Similarity) for evaluation.

4.3. Implementation details

The default visual inputs of the network consist of 16 frames of 224 × 224. The audio is resampled at 16 kHz and the audio spectrogram is generated using a 512 FFT and a 160 hop length setting. The initialization of the three-stream encoders is as discussed in Section 3.2. For the six audio-visual eye-tracking datasets, we adopt the same three-fold splitting strategy as STAViS [87]. We present the average results of each metric in the following experiments. We use the AdamW [58] optimizer with an initial learning rate of 1e-4, and the learning rate is reduced by a factor of 10 after 60 epochs. We train our model for a total of 120 epochs and monitor the model every 10 epochs. The batch size is set to 4. We evaluate our framework using various video backbones and larger visual inputs, which requires adjusting the batch size to accommodate a 12GB GPU. For smaller batch sizes, we replace the batchnorm layers in projectors and predictors with layernorm layers to avoid gradient explosion.

4.4. Comparison with state-of-the-art methods

We compare our methods with 5 state-of-the-art AVSP models on the six audio-visual eye-tracking datasets, as shown in Table 1. For our method, we offer 3D CNN-based and Transformer-based models with two types of resolutions (224 × 224 and 224 × 384) in Table 1. From Table 1, it can be observed that current AVSP models are trained on various visual resolutions and methods like STAViS are trained in a low resolution, which may hurt their performance. In the AVSP research field, most of the models have not released their codes, while AViNet and STAViS have released their codes. Thus, we retrain STAViS with 224 × 384 visual inputs and the results are denoted by * in Table 1. Compared to the original STAViS, STAViS* achieves improvements on the most of the evaluation metrics, like CC in AVAD from 0.6086 to 0.6539, CC in ETMD from 0.5690 to 0.5849, NSS in Coutrot2 from 5.28 to 5.64 and CC in Coutrot1 from 0.4722 to 0.4965, and obtains a similar performance in DIEM and SumMe. However, STAViS* still lags behind our methods on the six datasets.

For the 3D CNN-based model of ours, "Ours (S3D)" outperforms other AVSP models on the six audio-visual datasets apart from CASP-Net. Compared with CASP-Net, "Ours (S3D)" with 224 × 384 outperforms it on DIEM, AVAD, Coutrot1 and Coutrot2, and falls short on ETMD and SumMe. Our transformer-based model "Ours (MViTv2-S)" with 224 × 224 outperforms the above 3D CNN-based models on the most of evaluation metrics. With a higher resolution, our method achieves a 3.85% improvement on CC in Coutrot1, increasing from 0.5832 to 0.6057. Overall, our 3D CNN-based models show compelling performance with other AVSP models and the superior performance of the transformer-based method demonstrates that the strong ability to capture long-term dynamic saliency features benefits the AVSP task.

We also visualize some frame samples from DIEM, ETMD, SumMe, AVAD, Coutrot1 and Coutrot2, along with the corresponding saliency maps of the state-of-the-art AVSP models in Fig. 5. We can observe that in the ETMD column, our model pay attention to the shoe in man's hand while other models care more about the man's hand. In the DIEM column, STAViS and AViNet are influenced by the surround objects and there is significant difference from the groud truth. In the Coutrot2 column, our model correctly identifies the speaking man, while STAViS and AViNet mistakenly highlight the second person. Overall, our results

Table 2
Ablation studies of the image saliency features and audio-visual fusion on the test sets of DIEM, ETMD, SumMe, AVAD, Coutrot1 and Coutrot2 datasets. $\mathscr{E}_{\text{IS}}$ (ConvNeXt-T*) is not finetuned on image saliency datasets. $\mathscr{E}_{\text{IS}}$ (ConvNeXt-T†) directly provides saliency maps instead of image saliency features. The best results are **hightlighted**.

				DIEM [64]					ETMD [43, 86]					SumMe [31, 86]					AVAD [62]					Coutrot1 [20]					Coutrot2 [21]				
	$\mathcal{A}$	$\mathcal{L}$	$\mathscr{E}_{\text{IS}}$	CC↑	NSS↑	AUC↑	sAUC↑	SIM↑	CC↑	NSS↑	AUC↑	sAUC↑	SIM↑	CC↑	NSS↑	AUC↑	sAUC↑	SIM↑	CC↑	NSS↑	AUC↑	sAUC↑	SIM↑	CC↑	NSS↑	AUC↑	sAUC↑	SIM↑	CC↑	NSS↑	AUC↑	sAUC↑	SIM↑
Base		$\mathcal{L}_{\text{sal}}$		.6164	2.4591	.8981	.6843	.5071	.5661	2.9992	.9301	.7317	.4285	.4426	2.2392	.8924	.6688	.3448	.6690	3.7672	.9282	.6063	.5097	.5374	2.5495	.8866	.6048	.4291	.7482	6.0488	.9607	.7132	.5338
V1	✓	$\mathcal{L}_{total}$		.6308	2.5070	.9008	.6885	.5122	.5752	3.0506	.9321	.7345	.4337	.4571	2.3194	.8969	.6725	.3543	.6752	3.7774	.9298	.6085	.5145	.5411	2.5524	.8880	.6037	.4319	.7443	5.9660	.9610	.7094	.5262
V2		$\mathcal{L}_{\text{sal}}$	MobileNetV2 [77]	.6299	2.5231	.9018	.6863	.5182	.5787	3.0837	.9323	.7379	.4386	.4662	2.3865	.8997	.6814	.3605	.6751	3.7768	.9292	.6073	.5161	.5427	2.5908	.8882	.6052	.4328	.7629	6.1834	**.9616**	.7143	.5464
V3	✓	$\mathcal{L}_{\text{sal}}$	MobileNetV2 [77]	.6353	2.5247	.9014	.6895	.5172	.5839	3.1033	.9337	.7364	.4394	.4662	2.3597	.9005	.6785	.3566	.6639	3.6865	.9302	.6061	.5074	.5377	2.5240	.8879	.6035	.4288	.7502	6.1209	.9602	.7103	.5421
V4	✓	$\mathcal{L}_{total}$	MobileNetV2 [77]	.6340	2.5307	.9021	.6882	.5177	.5812	3.0923	.9331	.7371	.4394	.4733	2.4241	**.9010**	.6814	.3643	.6813	3.8041	.9304	.6088	.5194	.5417	2.5674	.8897	**.6078**	.4307	**.7693**	**6.2440**	.9611	**.7159**	.5516
V5	✓	$\mathcal{L}_{total}$	ConvNeXt-T [56]	**.6495**	**2.5934**	**.9057**	**.6942**	**.5265**	**.5919**	**3.1490**	**.9353**	**.7457**	.4415	**.4761**	**2.4511**	.8992	**.6822**	**.3672**	**.6875**	3.8238	**.9318**	**.6118**	**.5218**	.5403	2.5817	.8895	.6063	.4324	.7658	6.2071	.9612	.7151	**.5528**
V6	✓	$\mathcal{L}_{total}$	ConvNeXt-T* [56]	.6359	2.5382	.9023	.6879	.5159	.5880	3.1178	.9345	.7436	.4422	.4724	2.4045	.8993	.6803	.3590	.6747	3.7653	.9290	.6116	.5097	**.5483**	**2.6015**	**.8902**	.6056	.4323	.7447	6.0192	**.9616**	.7075	.5329
V7	✓	$\mathcal{L}_{total}$	ConvNeXt-T† [56]	.6487	2.5875	.9042	.6913	.5262	.5837	3.1149	.9335	.7391	**.4459**	.4682	2.3960	.8999	.6780	.3623	.6789	**3.8294**	.9300	.6087	.5217	.5401	2.5779	.8884	.6033	**.4357**	.7300	5.8826	.9603	.7091	.5231

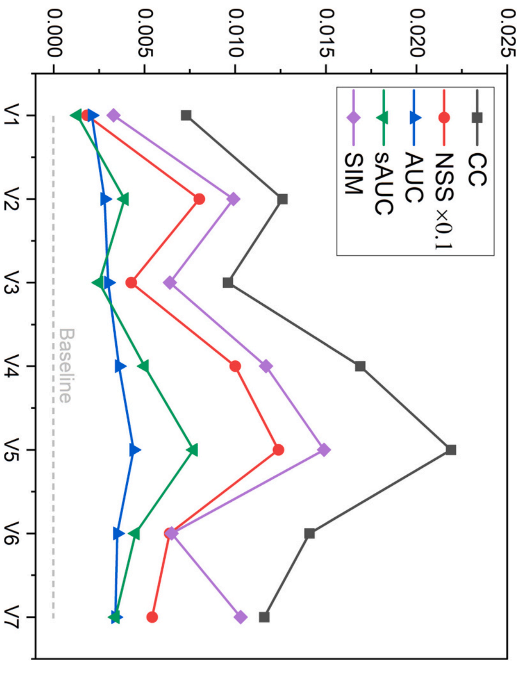

Fig. 6. Comparison of the seven variants (V1, V2, V3, V4, V5, V6 and V7) with the baseline model. The y-axis represents the difference in averaged metric scores of the six audio-visual eye-tracking datasets between the baseline and the compared variant. For a better display, the NSS scores are multiplied by 0.1.

Table 3
The performance of image saliency models with different encoders on the image saliency datasets, viz. the validation sets of SALICON and MIT1003. For the validation set of MIT1003, 103 images are selected for the evaluation. The best results are **highlighted**.

Encoder	SALICON [40]		MIT1003 [41]	
	CC↑	NSS↑	CC↑	NSS↑
MobileNetV2 [77]	0.8838	1.8646	0.7022	**0.8398**
ConvNeXt-T [56]	**0.9072**	**1.9281**	**0.7564**	0.8150

Table 4
Ablation studies of the audio-visual loss weight λ on the test sets. The average scores of each metric on the six datasets are given. The best results are **highlighted**.

λ	Average				
	CC↑	NSS↑	AUC↑	sAUC↑	SIM↑
0	0.6062	3.3865	0.9190	0.6707	0.4653
0.1	0.6072	3.3778	0.9189	0.6715	0.4636
0.5	0.6086	3.3859	0.9189	0.6716	0.4644
1	**0.6135**	**3.4438**	**0.9196**	**0.6732**	**0.4705**
3	0.6076	3.3948	0.9187	0.6703	0.4675

are closer to the ground truths.

4.5. Ablation study

In this section, we conduct ablation studies to evaluate the image saliency, audio and motion components in our framework. To evaluate the effectiveness of the image saliency encoder and the audio-visual fusion block, we adopt S3D and MobileNetV2 as the base settings for the motion encoder and image saliency encoder, respectively, for a convenient comparison. As shown in Table 2, we create seven variants. The baseline model is the visual-only model supervised by $\mathscr{L}_{\text{sal}}$. In V1, audio is introduced into the model and the variant is optimized by $\mathscr{L}_{total}$. In V2, the image saliency encoder is incorporated into our framework. For V3 and V4, the difference lies in whether the audio-visual symmetrized loss $\mathscr{L}_{\text{audio-visual}}$ is used or not. We further analyze the audio-visual method and audio-visual loss in Table 4 and Table 5. For V5, we replace

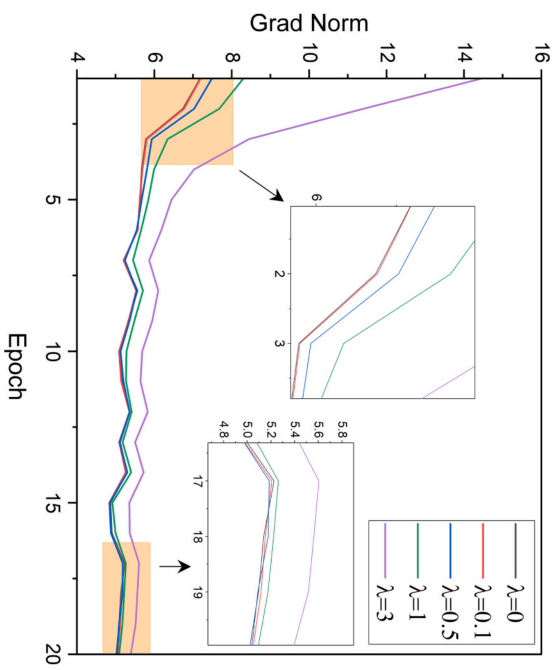

Fig. 7. The gradient norm curves during the first 20 epochs under different λ settings (batch size is 4). Zoom-in for a better view.

Table 5
Ablation studies of different audio-visual fusion methods on the test sets. The average scores of each metric on the six datasets are given. Avg refers to average pooling and Max refers to max pooling. C refers to the channel dimension and S refers to the spatial dimension. AVFB is our audio-visual fusion block. The best results are **highlighted**.

Method	Average				
	CC↑	NSS↑	AUC↑	sAUC↑	SIM↑
bilinear + Avg + C [87]	0.6045	3.3762	0.9181	0.6715	0.4626
bilinear + Max+S [36]	0.6013	3.3721	0.9175	0.6679	0.4648
AVIM [102]	0.6079	3.4181	0.9186	0.6707	0.4669
AVIM+CPC [102]	0.6096	3.4233	0.9189	0.6712	0.4675
AVFB	**0.6135**	**3.4438**	**0.9196**	**0.6732**	**0.4705**

the image saliency encoder in V4 with ConvNeXt-T. The V6 variant utilizes an image saliency encoder (ConvNeXt-T*) that is not trained on image saliency datasets. In V7, we directly fuse motion features with saliency maps generated by the image saliency encoder (ConvNeXt-T†) instead of using image saliency features. We calculate the average of the five metrics from the six audio-visual eye-tracking datasets and illustrate the difference in the averaged metric scores between the baseline and variants in Fig. 6. For motion features, we conduct experiments on the video backbones of various structures and the influence of temporal property to the AVSP task. The results are presented in Table 6, Fig. 8 and Fig. 9.

Ablation Study on the Image Saliency Features: As discussed in Section 3.3, it is beneficial for the performance of AVSP task by incorporating image saliency features. In Table 2, V2 variant shows improvements in every metrics on the six audio-visual eye-tracking datasets compared to the baseline model, which reflects that the image saliency encoder indeed can benefits the AVSP task. Additionally, it is predictable that the better performance of an image saliency model can result in more informative saliency information. Here, we use two different ISP models by replacing the image backbone. In Table 3, it can be observed that the ConvNeXt-T-based image saliency model achieves a better performance on the validation set of SALICON and MIT1003. Thus, we replace the MobileNetV2-based image saliency encoder with the stronger ConvNeXt-T-based one as the variant V5 in Table 2. Compared with V4, V5 variant obtains better performance and has a sharp growth in Fig. 6. When $\mathscr{E}_{\mathrm{IS}}$ is not equipped with image saliency-related knowledge, V6 suffers a performance degradation as shown in

Table 6
Ablation Study of the motion features on the test sets of DIEM, ETMD, SumMe, AVAD, Coutrot1 and Coutrot2 datasets. The size of visual inputs is 16 × 224 × 224. We also report the model parameters (Params) and multiplyâ€“accumulate operations (MACs) of our models. The frozen $\mathscr{E}_{\mathrm{A}}$ and $\mathscr{E}_{\mathrm{IS}}$ take up to 41.53 M in Params and 75.38 G in MACs. The best results are **highlighted**.

$\mathcal{E}_{\mathrm{V}}$	Params	MACs	DIEM [64]					ETMD [43, 86]					SumMe [31, 86]					AVAD [62]					Coutrot1 [20]					Coutrot2 [21]				
	(M)	(G)	CC↑	NSS↑	AUC↑	sAUC↑	SIM↑	CC↑	NSS↑	AUC↑	sAUC↑	SIM↑	CC↑	NSS↑	AUC↑	sAUC↑	SIM↑	CC↑	NSS↑	AUC↑	sAUC↑	SIM↑	CC↑	NSS↑	AUC↑	sAUC↑	SIM↑	CC↑	NSS↑	AUC↑	sAUC↑	SIM↑
S3D [101]	88	122	.6495	2.5934	.9057	.6942	.5265	.5919	3.1490	.9353	.7457	.4415	.4761	2.4511	.8992	.6822	.3672	.6875	3.8238	.9318	.6118	.5218	.5403	2.5817	.8895	.6063	.4324	.7658	6.2071	.9612	.7151	.5528
X3D-L [26]	84	121	.6466	2.5925	.9052	.6937	.5270	.5939	3.1639	.9356	.7443	.4493	.4809	2.4709	.9012	.6794	.3690	.6635	3.6519	.9304	.6023	.5046	.5527	2.6449	.8866	.6020	.4417	.7644	6.1221	.9616	.7101	.5620
SlowFast-4×16 [27]	114	127	**.6637**	**2.6573**	.9084	**.7025**	.5388	.6017	3.2132	.9361	.7475	.4566	.4772	2.4541	.9012	.6837	.3657	.6818	3.7654	.9322	.6105	.5186	.5725	2.7550	.8919	.6137	.4507	.7697	6.1224	**.9637**	.7162	.5634
MorphMLP-S [104]	126	172	.6575	2.6507	.9083	.6978	.5385	.6106	3.2705	.9391	.7483	**.4669**	.4916	2.5334	.9064	.6865	**.3804**	.6852	3.8464	.9315	.6106	.5217	.5682	2.7579	.8920	.6093	.4524	.7334	5.8983	.9614	.7042	.5463
VideoSwin-S [57]	129	188	.6603	2.6569	.9093	.6974	**.5405**	.6006	3.2009	.9374	.7450	.4584	.4849	2.4952	.9047	.6849	.3732	.6733	3.7515	.9298	.6082	.5198	.5609	2.7004	.8910	.6070	.4489	**.7763**	**6.3028**	.9631	.7179	.5621
Uniformer-B [52]	129	202	.6593	2.6557	**.9101**	.7001	.5392	.6110	**3.2781**	.9385	.7509	.4655	.4822	2.4946	.9028	.6887	.3731	.6863	3.8415	.9325	.6104	**.5275**	.5720	2.7688	.8922	.6099	.4529	.7599	6.2115	.9628	**.7185**	.5638
MViTv2-S [53]	114	170	.6592	2.6527	.9090	.6991	.5372	**.6147**	3.2679	**.9405**	**.7525**	.4633	**.4958**	**2.5535**	**.9075**	**.6942**	.3758	**.6927**	**3.8631**	**.9356**	**.6120**	.5247	**.5832**	**2.8249**	**.8946**	**.6162**	**.4594**	.7691	6.2281	.9622	.7115	**.5686**

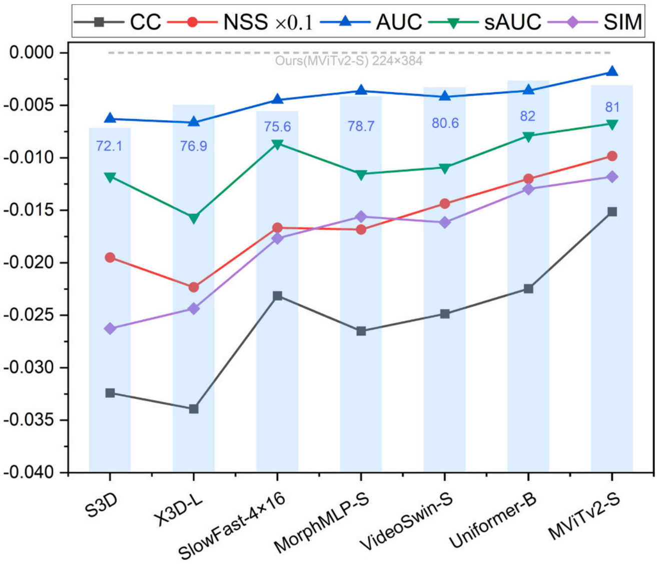

Fig. 8. Comparison of the motion encoders $\mathscr{E}_{\mathrm{V}}$ with various structures. The y-axis represents the difference in averaged metric scores of the six audio-visual eye-tracking datasets between "Ours (MViTv2-S)" with 224 × 384 visual inputs and the compared models. The bars in the figure represent the Top-1 accuracy of each video backbone on Kinetics 400. For a better display, the NSS scores are multiplied by 0.1.

Fig. 6. Additionally, V5 achieves better performance than V7, which shows that the fusion in feature level is better than direct fusion using image saliency maps, since the generated image saliency maps have no temporal information.

Ablation Study on the Audio-visual Fusion: As shown in Table 2 V1, the incorporation of audio features into saliency models leads to enhancements in all metrics for each dataset, in comparison to the baseline model. The consistent improvements on the six audio-visual eye-tracking datasets demonstrate the effectiveness of introducing the audio-visual fusion block. To testify the necessity of the symmetrized cosine similarity loss $\mathscr{L}_{\text{audio-visual}}$, we exclude it from the $\mathscr{L}_{total}$ in V3 variant. Compared to the V4 variant, the V3 variant achieves worse performance on SumMe, AVAD, Coutrot1 and Coutrot2. In particular, the CC scores of AVAD and Coutrot2 drop from 0.6813 to 0.6639 and from 0.7693 to 0.7502, respectively. In Fig. 6, we can observe a sharp decline on CC, NSS, sAUC and SIM metrics from the V4 variant to the V3 variant. It suggests that the symmetrized cosine similarity loss is a crucial component in our framework for AVSP. We also conduct experiments on the weight of the audio-visual loss λ. From Table 4, it can be observed that λ has an impact on the performance and too large or too small λ leads to a noticeable performance drop. λ heavily influences the learning direction of our model at the early optimization stage based on the gradient norm curves in Fig. 7. When λ is set to 3, the gradient norm value almost doubles as λ equals 0, making $\mathscr{L}_{\text{audio-visual}}$ and $\mathscr{L}_{\text{sal}}$ almost equally affect the optimization of our model at the early epoch. A performance drop can be observed in Table 4 compared with $\lambda = 1$. Our goal to generate saliency maps determines that the saliency loss $\mathscr{L}_{\text{sal}}$ plays a major role during the training process. Under an appropriate setting, the audio-visual loss $\mathscr{L}_{\text{audio-visual}}$ helps our AVSP model achieve a better performance.

Furthermore, we compare our audio-visual fusion method with those used in AVSP models [36,87,102]. Bilinear fusion operation [84], AVIM (Audio-Visual Interaction Module) and CPC (Consistency-aware Predictive Coding) are introduced to compare with the designed audio-visual fusion block. Those models with *bilinear* employ average pooling or max pooling along the temporal dimension, which removes the temporal information, before the later audio-visual fusion. STAViS performs *bilinear* along the channel dimension, while AViNet performs it along the spatial dimension. The experimental results are given in Table 5. The bilinear-based and AVIM-based models perform worse than our model. This demonstrates that our audio-visual fusion method achieves a better ability of audio-visual feature integration.

Ablation Study on the Motion Features: We conduct experiments on seven video backbones of 3D CNN, Transformer and MLP. From Table 6 and Fig. 8, we can observe that "MViTv2-S" achieves the best performance in terms of most evaluation metrics. Other Transformer-based models, such as VideoSwin and Uniformer, also achieve good performance. It suggests the long-range relationship captured by the Transformer is beneficial for AVSP. "MorphMLP-S" obtains worse performance than "SlowFast-4 × 16". In 3D CNN-based AVSP models, "SlowFast-4 × 16"achieves the highest performance and is comparable with Transformer-based models. SlowFast-4 × 16 falls short of X3D-L and MorphMLP-S on the Top-1 scores of Kinetics 400 in Fig. 8, but its variant still achieves better performance in AVSP. The reason is that SlowFast designs a fast branch to effectively capture dynamic features. Additionally, in Table 6, we report the model efficiency of the parameters (Params) and multiply-accumulate operations (MACs). The frozen $\mathscr{E}_{\mathrm{A}}$ and $\mathscr{E}_{\mathrm{IS}}$ take up to 41.53 M in Params and 75.38 G in MACs. It is observed that the greater performance of AVSP models requires more computing resources, as the parameters of "MViTv2-S" and "SlowFast-4 × 16"are similar but the "MViTv2-S" requires large computing burden.

To further observe the temporal relationship to the AVSP task, we infer "Ours (MViTv2-S)" in reverse order and the results are shown in Fig. 9. From Fig. 9, we can observe that the reverse inference has a more

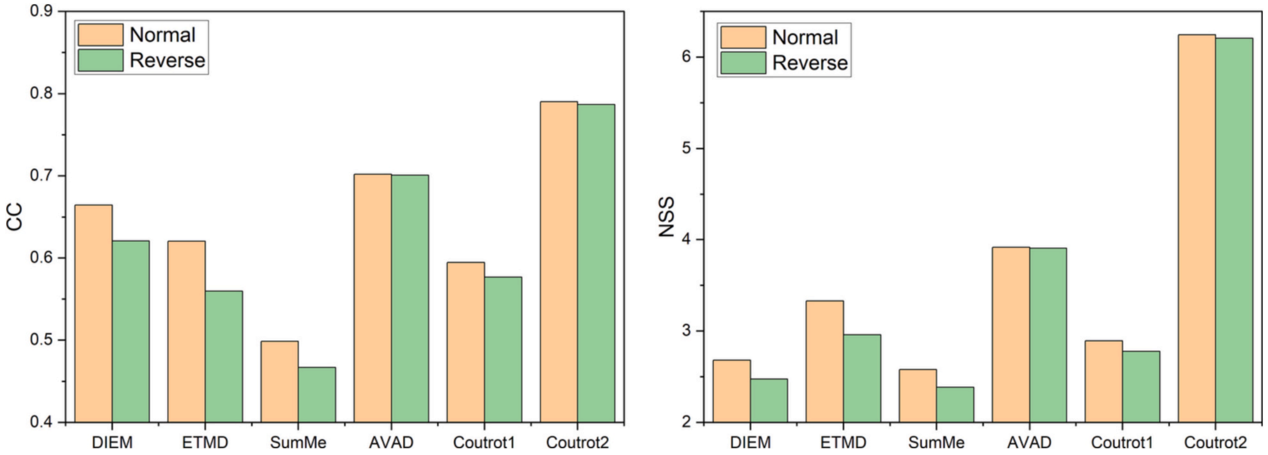

Fig. 9. The influence of the temporal order of video frames to the AVSP model's performance in the CC and the NSS metrics on the six audio-visual eye-tracking datasets.

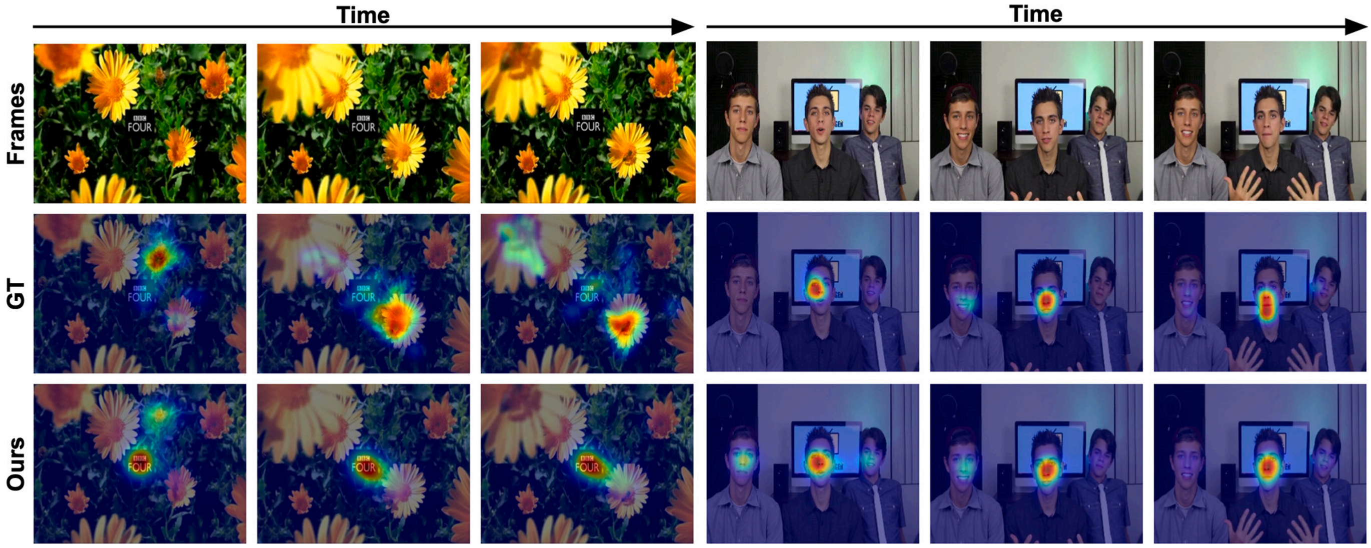

Fig. 10. Failure examples.

significant impact on datasets containing videos of longer duration like DIEM, SumMe and ETMD. However, it has a minimal impact on datasets containing videos of shorter duration like AVAD, Coutrot1 and Coutrot2. This behavior is highly related to the video content, indicating that the dataset bias about motion properties affects our AVSP models.

4.6. Failure cases and analysis

As indicated by the aforementioned experimental results, the proposed multi-sensory framework can achieve better performance on various audio-visual eye-tracking datasets. However, as shown in Fig. 10, the coexistence of static texts in the center and dynamic objects in natural scenes, such as the bees and texts within the frames, poses a challenge for our model in effectively capturing the inherent relationship between them. Our model focuses on the central texts while disregarding the presence of flying bees. The second example, shown in the fourth, fifth and sixth columns, portrays a scene in which three people seat together with the person in the middle talking. In such audio-visual scene, our model can generate good saliency maps as shown in the results of the second and third frames. However, our model exhibits a failure in accurately predicting the saliency map of the first frame. Since this frame is located at the early segment of the video clip, the collected eye-tracking data begins from the central point of the screen. Even though the audio-visual relationship has been established, it is important to consider the temporal order of a video from a holistic perspective, as it can also influence the eye movement behavior of humans.

5. Conclusion

In this paper, we present a multi-sensory framework for audio-visual saliency prediction by perceiving and integrating motion, audio and image saliency features. To obtain multi-sensory information, a three-stream encoder is utilized to extract audio, motion and image saliency features. In particular, we simplify the architecture of image saliency models and analyze the operating mechanism of image saliency features in our framework. In order to tackle the issue of the weak ability to capture long-term motion features in 3D convolutions, we introduce various Transformer-based and MLP-based video backbones to extract dynamic saliency features for AVSP. To learn joint audio-visual representations in a self-supervised manner, an audio-visual fusion block is designed to enhance the audio-visual correspondence features with the supervision of a symmetrized cosine similarity loss. A multi-stage decoder is used to integrate audio-visual and multi-stage motion features in order to generate the final saliency map. We conduct comprehensive experiments on the image saliency, audio and motion features in our framework. Experimental results on the six audio-visual eye-tracking datasets demonstrate that our models achieve impressive performance compared to other state-of-the-art AVSP models.

CRediT authorship contribution statement

Jiawei Xie: Conceptualization, Investigation, Methodology, Software, Visualization, Writing – original draft. **Zhi Liu:** Conceptualization, Funding acquisition, Project administration, Resources, Supervision, Writing – review & editing. **Gongyang Li:** Funding acquisition, Writing – review & editing. **Yingjie Song:** Formal analysis, Validation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 62171269, and in part by the China Postdoctoral Science Foundation under Grant 2022M722037.

References

[1] T. Afouras, J.S. Chung, A. Senior, O. Vinyals, A. Zisserman, Deep audio-visual speech recognition, IEEE Trans. Pattern Anal. Mach. Intell. 44 (2018) 8717–8727, https://doi.org/10.1109/TPAMI.2018.2889052.

[2] T. Afouras, A. Owens, J.S. Chung, A. Zisserman, Self-supervised learning of audio-visual objects from video, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16, Springer, 2020, pp. 208–224, https://doi.org/10.1007/978-3-030-58523-5_13.

[3] H. Alwassel, D. Mahajan, B. Korbar, L. Torresani, B. Ghanem, D. Tran, Self-supervised learning by cross-modal audio-video clustering, Adv. Neural Inf. Proces. Syst. 33 (2020) 9758–9770.

[4] R. Arandjelovic, A. Zisserman, Look, listen and learn, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 609–617, https://doi.org/10.1109/ICCV.2017.73.

[5] B. Aydemir, L. Hoffstetter, T. Zhang, M. Salzmann, S. Süsstrunk, Tempsal-uncovering temporal information for deep saliency prediction, in: Proceedings of

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 6461–6470, https://doi.org/10.1109/CVPR52729.2023.00625.

[6] Y. Aytar, C. Vondrick, A. Torralba, Soundnet: Learning sound representations from unlabeled video, in: Advances in Neural Information Processing Systems, 2016, p. 29.

[7] N. Bruce, J. Tsotsos, Saliency based on information maximization, in: Advances in Neural Information Processing Systems, 2005, p. 18.

[8] M. Cerf, E.P. Frady, C. Koch, Faces and text attract gaze independent of the task: experimental data and computer model, J. Vis. 9 (2009) 10, https://doi.org/10.1167/9.12.10.

[9] Q. Chang, S. Zhu, Temporal-spatial feature pyramid for video saliency detection, 2021 arXiv preprint arXiv:2105.04213.

[10] F.Y. Chao, C. Ozcinar, C. Wang, E. Zerman, L. Zhang, W. Hamidouche, O. Deforges, A. Smolic, Audio-visual perception of omnidirectional video for virtual reality applications, in: 2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), IEEE, 2020, pp. 1–6, https://doi.org/10.1109/ICMEW46912.2020.9105956.

[11] C. Chen, M. Song, W. Song, L. Guo, M. Jian, A comprehensive survey on video saliency detection with auditory information: the audio-visual consistency perceptual is the key!, IEEE Trans. Circuits Syst. Video Technol. 33 (2022) 457–477, https://doi.org/10.1109/TCSVT.2022.3203421.

[12] H. Chen, W. Xie, A. Vedaldi, A. Zisserman, Vggsound: A large-scale audio-visual dataset, in: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2020, pp. 721–725, https://doi.org/10.1109/ICASSP40776.2020.9053174.

[13] J. Chen, H. Song, K. Zhang, B. Liu, Q. Liu, Video saliency prediction using enhanced spatiotemporal alignment network, Pattern Recogn. 109 (2021) 107615, https://doi.org/10.1016/j.patcog.2020.107615.

[14] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: International Conference on Machine Learning, PMLR, 2020, pp. 1597–1607.

[15] X. Chen, K. He, Exploring simple siamese representation learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 15750–15758, https://doi.org/10.1109/CVPR46437.2021.01549.

[16] Y. Cheng, R. Wang, Z. Pan, R. Feng, Y. Zhang, Look, listen, and attend: Co-attention network for self-supervised audio-visual representation learning, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 3884–3892, https://doi.org/10.1145/3394171.3413869.

[17] M. Cornia, L. Baraldi, G. Serra, R. Cucchiara, A deep multi-level network for saliency prediction, in: 2016 23rd International Conference on Pattern Recognition (ICPR), IEEE, 2016, pp. 3488–3493.

[18] M. Cornia, L. Baraldi, G. Serra, R. Cucchiara, Predicting human eye fixations via an lstm-based saliency attentive model, IEEE Trans. Image Process. 27 (2018) 5142–5154, https://doi.org/10.1109/TIP.2018.2851672.

[19] A. Coutrot, N. Guyader, An audiovisual attention model for natural conversation scenes, in: 2014 IEEE International Conference on Image Processing (ICIP), IEEE, 2014, pp. 1100–1104, https://doi.org/10.1109/ICIP.2014.7025219.

[20] A. Coutrot, N. Guyader, How saliency, faces, and sound influence gaze in dynamic social scenes, J. Vis. 14 (2014) 5, https://doi.org/10.1167/14.8.5.

[21] A. Coutrot, N. Guyader, Multimodal saliency models for videos, in: From Human Attention to Computational Attention: A Multidisciplinary Approach, 2016, pp. 291–304, https://doi.org/10.1007/978-1-4939-3435-5_16.

[22] G. Ding, N. İmamoğlu, A. Caglayan, M. Murakawa, R. Nakamura, Salfbnet: learning pseudo-saliency distribution via feedback convolutional networks, Image Vis. Comput. 120 (2022) 104395, https://doi.org/10.1016/j.imavis.2022.104395.

[23] R. Droste, J. Jiao, J.A. Noble, Unified image and video saliency modeling, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16, Springer, 2020, pp. 419–435, https://doi.org/10.1007/978-3-030-58558-7_25.

[24] H. Duan, Z. Liu, W. Wei, T. Zhang, J. Wang, L. Xu, H. Liu, T. Chen, Atypical salient regions enhancement network for visual saliency prediction of individuals with autism spectrum disorder, Signal Process. Image Commun. 115 (2023) 116968, https://doi.org/10.1016/j.image.2023.116968.

[25] E. Erdem, A. Erdem, Visual saliency estimation by nonlinearly integrating features using region covariances, J. Vis. 13 (2013) 11, https://doi.org/10.1167/13.4.11.

[26] C. Feichtenhofer, X3d: Expanding architectures for efficient video recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 203–213, https://doi.org/10.1109/CVPR42600.2020.00028.

[27] C. Feichtenhofer, H. Fan, J. Malik, K. He, Slowfast networks for video recognition, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 6202–6211.

[28] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K.V. Alwala, A. Joulin, I. Misra, Imagebind: One embedding space to bind them all, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 15180–15190, https://doi.org/10.1109/CVPR52729.2023.01457.

[29] Y. Gong, A. Rouditchenko, A.H. Liu, D. Harwath, L. Karlinsky, H. Kuehne, J. R. Glass, Contrastive audio-visual masked autoencoder, in: The Eleventh International Conference on Learning Representations, 2022.

[30] J.B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, et al., Bootstrap your own latent-a new approach to self-supervised learning, Adv. Neural Inf. Proces. Syst. 33 (2020) 21271–21284.

[31] M. Gygli, H. Grabner, H. Riemenschneider, L.V. Gool, Creating summaries from user videos, in: European Conference on Computer Vision, 2014, https://doi.org/10.1007/978-3-319-10584-0_33.

[32] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[33] S. Hershey, S. Chaudhuri, D.P. Ellis, J.F. Gemmeke, A. Jansen, R.C. Moore, M. Plakal, D. Platt, R.A. Saurous, B. Seybold, et al., Cnn architectures for large-scale audio classification, in: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2017, pp. 131–135.

[34] P.Y. Huang, H. Xu, J. Li, A. Baevski, M. Auli, W. Galuba, F. Metze, C. Feichtenhofer, Masked autoencoders that listen, in: Advances in Neural Information Processing Systems 35, 2022, pp. 28708–28720.

[35] L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis, IEEE Trans. Pattern Anal. Mach. Intell. 20 (1998) 1254–1259, https://doi.org/10.1109/34.730558.

[36] S. Jain, P. Yarlagadda, S. Jyoti, S. Karthik, S. Ramanathan, V. Gandhi, Vinet: Pushing the limits of visual modality for audio-visual saliency prediction, in: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2020, pp. 3520–3527, https://doi.org/10.1109/IROS51168.2021.9635989.

[37] S. Jia, N.D. Bruce, Eml-net: An expandable multi-layer network for saliency prediction, in: Image and Vision Computing 95, 2020 103887, https://doi.org/10.1016/j.imavis.2020.103887.

[38] L. Jiang, M. Xu, T. Liu, M. Qiao, Z. Wang, Deepvs: A deep learning based video saliency prediction approach, in: European Conference on Computer Vision, 2018, https://doi.org/10.1007/978-3-030-01264-9_37.

[39] L. Jiang, M. Xu, Z. Wang, Predicting video saliency with object-to-motion cnn and two-layer convolutional lstm, 2017 arXiv preprint arXiv:1709.06316.

[40] M. Jiang, S. Huang, J. Duan, Q. Zhao, Salicon: Saliency in context, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1072–1080, https://doi.org/10.1109/CVPR.2015.7298710.

[41] T. Judd, K. Ehinger, F. Durand, A. Torralba, Learning to predict where humans look, in: 2009 IEEE 12th International Conference on Computer Vision, IEEE, 2009, pp. 2106–2113, https://doi.org/10.1109/ICCV.2009.5459462.

[42] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, et al., The kinetics human action video dataset, 2017 arXiv preprint arXiv:1705.06950.

[43] P. Koutras, P. Maragos, A perceptually based spatio-temporal computational framework for visual saliency estimation, Signal Process. Image Commun. 38 (2015) 15–31, https://doi.org/10.1016/j.image.2015.08.004.

[44] M. Kümmerer, L. Theis, M. Bethge, Deep gaze i: Boosting saliency prediction with feature maps trained on imagenet, 2014 arXiv preprint arXiv:1411.1045.

[45] M. Kümmerer, T. Wallis, M. Bethge, Deepgaze ii: predicting fixations from deep features over time and tasks, J. Vis. 17 (2017) 1147, https://doi.org/10.1167/17.10.1147.

[46] M. Kümmerer, T.S. Wallis, L.A. Gatys, M. Bethge, Understanding low-and high-level contributions to fixation prediction, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 4789–4798, https://doi.org/10.1109/ICCV.2017.513.

[47] Q. Lai, W. Wang, H. Sun, J. Shen, Video saliency prediction using spatiotemporal residual attentive networks, IEEE Trans. Image Process. 29 (2020) 1113–1126, https://doi.org/10.1109/TIP.2019.2936112.

[48] G. Li, Z. Bai, Z. Liu, Texture-semantic collaboration network for orsi salient object detection, in: IEEE Transactions on Circuits and Systems II: Express Briefs, 2023, https://doi.org/10.1109/TCSII.2023.3333436.

[49] G. Li, Z. Bai, Z. Liu, X. Zhang, H. Ling, Salient object detection in optical remote sensing images driven by transformer, IEEE Trans. Image Process. 32 (2023) 5257–5269, https://doi.org/10.1109/TIP.2023.3314285.

[50] G. Li, Z. Liu, W. Lin, H. Ling, Multi-content complementation network for salient object detection in optical remote sensing images, IEEE Trans. Geosci. Remote Sens. 60 (2022) 1–13, https://doi.org/10.1109/TGRS.2021.3131221.

[51] G. Li, Z. Liu, D. Zeng, W. Lin, H. Ling, Adjacent context coordination network for salient object detection in optical remote sensing images, IEEE Trans. Cybernet. 53 (2023) 526–538, https://doi.org/10.1109/TCYB.2022.3162945.

[52] K. Li, Y. Wang, J. Zhang, P. Gao, G. Song, Y. Liu, H. Li, Y. Qiao, Uniformer: unifying convolution and self-attention for visual recognition, IEEE Trans. Pattern Anal. Mach. Intell. (2023), https://doi.org/10.1109/TPAMI.2023.3282631.

[53] Y. Li, C.Y. Wu, H. Fan, K. Mangalam, B. Xiong, J. Malik, C. Feichtenhofer, Mvitv2: Improved multiscale vision transformers for classification and detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 4804–4814, https://doi.org/10.1109/CVPR52688.2022.00476.

[54] A. Linardos, M. Kümmerer, O. Press, M. Bethge, Deepgaze iie: Calibrated prediction in and out-of-domain for state-of-the-art saliency modeling, in: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 12899–12908, https://doi.org/10.1109/ICCV48922.2021.01268.

[55] N. Liu, J. Han, D. Zhang, S. Wen, T. Liu, Predicting eye fixations using convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, https://doi.org/10.1109/CVPR.2015.7298633.

[56] Z. Liu, H. Mao, C.Y. Wu, C. Feichtenhofer, T. Darrell, S. Xie, A convnet for the 2020s, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 11976–11986, https://doi.org/10.1109/CVPR52688.2022.01167.

[57] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, H. Hu, Video swin transformer, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern

Recognition, 2022, pp. 3202–3211, https://doi.org/10.1109/CVPR52688.2022.00320.

[58] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: International Conference on Learning Representations, 2018.

[59] C. Ma, H. Sun, Y. Rao, J. Zhou, J. Lu, Video saliency forecasting transformer, IEEE Trans. Circuits Syst. Video Technol. 32 (2022) 6850–6862, https://doi.org/10.1109/TCSVT.2022.3172971.

[60] S. Mathe, C. Sminchisescu, Actions in the eye: dynamic gaze datasets and learnt saliency models for visual recognition, IEEE Trans. Pattern Anal. Mach. Intell. 37 (2014) 1408–1424, https://doi.org/10.1109/TPAMI.2014.2366154.

[61] K. Min, J.J. Corso, Tased-net: Temporally-aggregating spatial encoder-decoder network for video saliency detection, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2394–2403, https://doi.org/10.1109/ICCV.2019.00248.

[62] X. Min, G. Zhai, K. Gu, X. Yang, Fixation prediction through multimodal analysis, ACM Trans. Multimed. Comput. Commun. Appl. (TOMM) 13 (2016) 1–23, https://doi.org/10.1145/2996463.

[63] X. Min, G. Zhai, J. Zhou, X.P. Zhang, X. Yang, X. Guan, A multimodal saliency model for videos with high audio-visual correspondence, IEEE Trans. Image Process. 29 (2020) 3805–3819, https://doi.org/10.1109/TIP.2020.2966082.

[64] P.K. Mital, T.J. Smith, R.L. Hill, J.M. Henderson, Clustering of gaze during dynamic scene viewing is predicted by motion, Cogn. Comput. 3 (2011) 5–24, https://doi.org/10.1007/s12559-010-9074-z.

[65] A. Nagrani, S. Yang, A. Arnab, A. Jansen, C. Schmid, C. Sun, Attention bottlenecks for multimodal fusion, in: Advances in Neural Information Processing Systems 34, 2021, pp. 14200–14213.

[66] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, K. Kashino, Byol for audio: Exploring pre-trained general-purpose audio representations, in: IEEE/ACM Transactions on Audio, Speech, and Language Processing 31, 2022, pp. 137–151, https://doi.org/10.1109/TASLP.2022.3221007.

[67] H. Ning, B. Zhao, Z. Hu, L. He, E. Pei, Audio–visual collaborative representation learning for dynamic saliency prediction, Knowl.-Based Syst. 256 (2022) 109675, https://doi.org/10.1016/j.knosys.2022.109675.

[68] A. Owens, A.A. Efros, Audio-visual scene analysis with self-supervised multisensory features, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 631–648, https://doi.org/10.1007/978-3-030-01231-1_39.

[69] J. Pan, C.C. Ferrer, K. McGuinness, N.E. O'Connor, J. Torres, E. Sayrol, X. Giro-i Nieto, Salgan: Visual saliency prediction with generative adversarial networks, 2017 arXiv preprint arXiv:1701.01081.

[70] J. Pan, E. Sayrol, X. Giro-i Nieto, K. McGuinness, N.E. O'Connor, Shallow and deep convolutional networks for saliency prediction, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 598–606.

[71] Y. Pang, X. Zhao, L. Zhang, H. Lu, Multi-scale interactive network for salient object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 9413–9422.

[72] D.R. Perrott, K. Saberi, K. Brown, T.Z. Strybel, Auditory psychomotor coordination and visual search performance, Percept. Psychophys. 48 (1990) 214–226, https://doi.org/10.3758/BF03211521.

[73] E. Prashnani, O. Gallo, J. Kim, J.B. Spjut, P. Sen, I. Frosio, Noise-aware video saliency prediction, in: British Machine Vision Conference, 2021.

[74] A. Recasens, J. Lin, J. Carreira, D. Jaegle, L. Wang, J.B. Alayrac, P. Luc, A. Miech, L. Smaira, R. Hemsley, et al., Zorro: the masked multimodal transformer, 2023 arXiv preprint arXiv:2301.09595.

[75] A. Recasens, P. Luc, J.B. Alayrac, L. Wang, F. Strub, C. Tallec, M. Malinowski, V. Pătrăucean, F. Altché, M. Valko, et al., Broaden your views for self-supervised video learning, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 1255–1265, https://doi.org/10.1109/ICCV48922.2021.00129.

[76] N. Reddy, S. Jain, P. Yarlagadda, V. Gandhi, Tidying deep saliency prediction architectures, in: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2020, pp. 10241–10247, https://doi.org/10.1109/IROS45743.2020.9341574.

[77] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.C. Chen, Mobilenetv2: inverted residuals and linear bottlenecks, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (2018) 4510–4520, https://doi.org/10.1109/CVPR.2018.00474.

[78] P. Sarkar, A. Etemad, Self-supervised audio-visual representation learning with relaxed cross-modal synchronicity, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2023, pp. 9723–9732, https://doi.org/10.1609/aaai.v37i8.26162.

[79] N. Shvetsova, B. Chen, A. Rouditchenko, S. Thomas, B. Kingsbury, R.S. Feris, D. Harwath, J. Glass, H. Kuehne, Everything at once-multi-modal fusion transformer for video retrieval, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 20020–20029, https://doi.org/10.1109/CVPR52688.2022.01939.

[80] G. Song, D. Pellerin, L. Granjon, Different types of sounds influence gaze differently in videos, J. Eye Mov. Res. 6 (2013) 1–13, https://doi.org/10.16910/jemr.6.4.1.

[81] Y. Song, Z. Liu, G. Li, D. Zeng, T.H. Zhang, L. Xu, J. Wang, Rinet: relative importance-aware network for fixation prediction, IEEE Trans. Multimed. 25 (2023) 9263–9277, https://doi.org/10.1109/TMM.2023.3249481.

[82] B.W. Tatler, M.M. Hayhoe, M.F. Land, D.H. Ballard, Eye guidance in natural vision: reinterpreting salience, J. Vis. 11 (2011) 5, https://doi.org/10.1167/11.5.5.

[83] H.R. Tavakoli, A. Borji, E. Rahtu, J. Kannala, Dave: A deep audio-visual embedding for dynamic saliency prediction, 2019. ArXiv abs/1905.10693.

[84] J.B. Tenenbaum, W.T. Freeman, Separating style and content with bilinear models, Neural Comput. 12 (2000) 1247–1283, https://doi.org/10.1162/089976600300015349.

[85] A. Torralba, A. Oliva, M.S. Castelhano, J.M. Henderson, Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search, Psychol. Rev. 113 (2006) 766, https://doi.org/10.1037/0033-295X.113.4.766.

[86] A. Tsiami, P. Koutras, A. Katsamanis, A. Vatakis, P. Maragos, A behaviorally inspired fusion approach for computational audiovisual saliency modeling, Signal Process. Image Commun. 76 (2019) 186–200, https://doi.org/10.1016/j.image.2019.05.001.

[87] A. Tsiami, P. Koutras, P. Maragos, Stavis: Spatio-temporal audiovisual saliency network, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4765–4775, https://doi.org/10.1109/CVPR42600.2020.00482.

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

[89] E. Vig, M. Dorr, D. Cox, Large-scale optimization of hierarchical features for saliency prediction in natural images, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 2798–2805.

[90] J. Vroomen, B.D. Gelder, Sound enhances visual perception: cross-modal effects of auditory organization on vision, J. Exp. Psychol. Hum. Percept. Perform. 26 (2000) 1583, https://doi.org/10.1037//0096-1523.26.5.1583.

[91] G. Wang, C. Chen, D.P. Fan, A. Hao, H. Qin, From semantic categories to fixations: A novel weakly-supervised visual-auditory saliency detection approach, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 15119–15128, https://doi.org/10.1109/CVPR46437.2021.01487.

[92] W. Wang, J. Shen, Deep visual attention prediction, IEEE Trans. Image Process. 27 (2017) 2368–2378, https://doi.org/10.1109/TIP.2017.2787612.

[93] W. Wang, J. Shen, X. Dong, A. Borji, Salient object detection driven by fixation prediction, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 1711–1720, https://doi.org/10.1109/CVPR.2018.00184.

[94] W. Wang, J. Shen, X. Dong, A. Borji, R. Yang, Inferring salient objects from human fixations, IEEE Trans. Pattern Anal. Mach. Intell. 42 (2020) 1913–1927, https://doi.org/10.1109/TPAMI.2019.2905607.

[95] W. Wang, J. Shen, F. Guo, M.M. Cheng, A. Borji, Revisiting video saliency: A large-scale benchmark and a new model, in: IEEE Conference on Computer Vision and Pattern Recognition, 2018, https://doi.org/10.1109/CVPR.2018.00514.

[96] Y. Wang, Z. Liu, Y. Xia, C. Zhu, D. Zhao, Spatiotemporal module for video saliency prediction based on self-attention, in: Image and Vision Computing 112, 2021 104216, https://doi.org/10.1016/j.imavis.2021.104216.

[97] Z. Wang, Z. Liu, G. Li, Y. Wang, T.H. Zhang, L. Xu, J. Wang, Spatio-temporal self-attention network for video saliency prediction, IEEE Trans. Multimed. 25 (2021) 1161–1174, https://doi.org/10.1109/TMM.2021.3139743.

[98] Z. Wang, Z. Liu, W. Wei, H. Duan, Saled: Saliency prediction with a pithy encoder-decoder architecture sensing local and global information, in: Image and Vision Computing 109, 2021 104149, https://doi.org/10.1016/j.imavis.2021.104149.

[99] X. Wu, Z. Wu, J. Zhang, L. Ju, S. Wang, Salsac: A video saliency prediction model with shuffled attentions and correlation-based convlstm, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2020, pp. 12410–12417, https://doi.org/10.1609/aaai.v34i07.6927.

[100] J. Xie, Z. Liu, G. Li, X. Lu, T. Chen, Global semantic-guided network for saliency prediction, Knowl.-Based Syst. 284 (2024) 111279, https://doi.org/10.1016/j.knosys.2023.111279.

[101] S. Xie, C. Sun, J. Huang, Z. Tu, K. Murphy, Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 305–321, https://doi.org/10.1007/978-3-030-01267-0_19.

[102] J. Xiong, G. Wang, P. Zhang, W. Huang, Y. Zha, G. Zhai, Casp-net: Rethinking video saliency prediction from an audio-visual consistency perceptual perspective, in: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6441–6450, https://doi.org/10.1109/CVPR52729.2023.00623.

[103] S. Yang, G. Lin, Q. Jiang, W. Lin, A dilated inception network for visual saliency prediction, IEEE Trans. Multimed. 22 (2019) 2163–2176, https://doi.org/10.1109/TMM.2019.2947352.

[104] D.J. Zhang, K. Li, Y. Wang, Y. Chen, S. Chandra, Y. Qiao, L. Liu, M.Z. Shou, Morphmlp: An efficient mlp-like backbone for spatial-temporal representation learning, in: European Conference on Computer Vision 230–248, Springer, 2022, https://doi.org/10.1007/978-3-031-19833-5_14.

[105] Y. Zhang, T. Zhang, C. Wu, Y. Zheng, Accurate video saliency prediction via hierarchical fusion and temporal recurrence, Image Vis. Comput. 136 (2023) 104744, https://doi.org/10.1016/j.imavis.2023.104744.

[106] J. Zhou, J. Wang, J. Zhang, W. Sun, J. Zhang, S. Birchfield, D. Guo, L. Kong, M. Wang, Y. Zhong, Audio–visual segmentation, in: European Conference on Computer Vision, Springer, 2022, pp. 386–403, https://doi.org/10.1007/978-3-031-19836-6_22.

[107] X. Zhou, S. Wu, R. Shi, B. Zheng, S. Wang, H. Yin, J. Zhang, C. Yan, Transformer-based multi-scale feature integration network for video saliency prediction, IEEE Trans. Circuits Syst. Video Technol. 1–1 (2023), https://doi.org/10.1109/TCSVT.2023.3278410.

[108] D. Zhu, X. Shao, Q. Zhou, X. Min, G. Zhai, X. Yang, A novel lightweight audio-visual saliency model for videos, ACM Trans. Multimed. Comput. Commun. Appl. 19 (2023) 1–22, https://doi.org/10.1145/3576857.