Audio-Visual Collaborative Representation Learning for Dynamic Saliency Prediction

Hailong Ning[a], Bin Zhao[b,*], Zhanxuan Hu[a], Lang He[a], Ercheng Pei[a]

[a]*School of Computer Science and Technology, Xi'an University of Posts and Telecommunications, Xi'an 710121, P. R. China.*
[b]*School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, P. R. China.*

Abstract

The Dynamic Saliency Prediction (DSP) task simulates the human selective attention mechanism to perceive the dynamic scene, which is significant and imperative in many vision tasks. Most of existing methods only consider visual cues, while neglect the accompanied audio information, which can provide complementary information for the scene understanding. In fact, there exists a strong relation between auditory and visual cues, and humans generally perceive the surrounding scene by collaboratively sensing these cues. Motivated by this, an audio-visual collaborative representation learning method is proposed for the DSP task, which explores the audio modality to better predict the dynamic saliency map by assisting vision modality. The proposed method consists of three parts: 1) audio-visual encoding, 2) audio-visual location, and 3) collaborative integration parts. Firstly, a refined SoundNet architecture is adopted to encode audio modality for obtaining corresponding features, and a modified 3D ResNet-50 architecture is employed to learn visual features, containing both spatial location and temporal motion information. Secondly, an audio-visual location part is devised to locate the sound source in the visual scene by learning the correspondence between audio-visual information. Thirdly, a collaborative integration part is devised to adaptively aggregate audio-visual information and center-bias prior to generate the final saliency map. Extensive experiments are conducted on six challenging audiovisual eye-tracking datasets, including DIEM, AVAD, Coutrot1, Coutrot2, SumMe, and ETMD,

---

*Corresponding author

which shows significant superiority over state-of-the-art DSP models.

1. Introduction

Saliency prediction task aims to automatically predict the most prominent area in the scene by simulating the human selective attention mechanism, which provides an alternative for obtaining the most valuable information from massive data. The task has served an important research topic in the field of computer vision, and can be of great applications in many fields, such as scene understanding [1, 2], object detection [3, 4], object tracking [5], image quality evaluation [6], automatic contrast enhancement [7], and video compression [8].

In the field of computer vision, the saliency prediction task draws increasing attention, and lots of methods have been proposed in recent years [9–13]. According to the data type, the existing methods can be divided into two aspects, including Static Saliency Prediction (SSP) methods and Dynamic Saliency Prediction (DSP) methods. The SSP methods aim to leverage the low-level contrast information and high-level semantic information of images to achieve the prediction of prominent area in the scene [14]. As an early exploration of the saliency prediction task, Itti *et al.* [9] imitate the human bottom-up visual selective attention process to extract the low-level features of images, predict the corresponding saliency map. With the popularity of deep learning [15–17], a large number of researchers conduct the saliency prediction task by mining high-level semantic information in images [13, 18–20]. For example, Wang *et al.* [13] obtain the hierarchically saliency information by extracting multi-scale features of images. The DSP methods focus on applying the spatio-temporal structure information in the video for the prediction of prominent area in the scene. For example, Zhang *et al.* [21] design a spatial-temporal DSP model for learning spatial features with a static network and temporal features with a dynamic network. In view of practical applications, this paper aims to investigate the saliency prediction for the dynamic video.

1.1. Motivation and Overview

Generally, video data naturally includes two modalities, *i.e.*, audio and vision. They represent the scene content in the video from different aspects, and can complement to help the viewer better understand the video content. Recently, lots of multi-modal studies based on audio-visual data have shown that audios can significantly promote the understanding of the scene [22–25]. For example, Hu *et al.* [23] proposed a cross-task transfer learning model for scene classification based on audio-visual data, which shows the benefits of audio-visual analysis compared with single-modality analysis.

However, most researchers in DSP have not fully realized the potential contribution of audio information to the performance. They predict the dynamic saliency map by only mining the information in the visual modality itself, while ignoring the latent effect of the accompanied audio information.

It is well known that human attention is naturally influenced by audio-visual stimuli rather than only auditory or visual stimuli in isolation. Inspired by this, we argue that audio information can assist the vision modality to better predict the saliency map in this paper. Specifically, humans practically pay more attention to the sounding object in the video. With the assistance of audio-visual information, it is easily to locate more salient sounding object based on audio-visual consistency learning. In addition, human instinctively pays more attention to some other cues, such as the moving, center, and high-level semantic objects. However, there are often some inconsistencies among the influences for DSP by multiple cues. For example, 1) when existing multiple moving objects in a scene, how to locate the more prominent sounding objects; 2) When the original audio is replaced with background sound, or when the sounding objects is not in the filed of the view, how to locate the salient objects; 3) As for the effect by multiple cues, *i.e.* movement, sound, and center-prior, how to better integrate them. These inconsistencies will bring some interference and are the main challenges for the audio-visual DSP.

Based on the above opportunities and challenges, an audio-visual collaborative representation learning method is proposed for the DSP task. The proposed method can be purposely leveraged to predict the dynamic saliency map in videos by collaboratively integrating the audio-visual cues. Concretely, the proposed method is composed

3

of three main parts: 1) audio-visual encoding, 2) audio-visual location, and 3) collaborative integration parts. Firstly, the audio-visual encoding part consists of two branches, which are responsible for learning audio and visual features with spatial location and temporal motion information, respectively. Specifically, the audio branch adopts a refined SoundNet architecture [26], and the visual branch employs a modified 3D ResNet-50 architecture [27]. Secondly, with the proposed audio-visual location part, the learned audio-visual features are jointly to locate the sound source in the visual scene by learning the correspondence between audio-visual information. Thirdly, the collaborative integration part is introduced to adaptively aggregate multi-cues audio-visual information, so as to generate the final dynamic saliency map. Here, the multi-cues information indicates the localized sound source information, spatio-temporal visual information (motion, high-level semantics, *etc.*), as well as center-bias prior.

1.2. Contributions

Generally, the main contributions of this paper are threefold:

- An audio-visual collaborative representation learning method is proposed, which can comprehensively consider the influence of audio-visual cues on dynamic saliency prediction, and realize the role of auxiliary enhancement.

- An audio-visual location part is devised to learn the correspondence between audio and visual modalities, so as to predict the sounding objects in the scene and reduce interference.

- A collaborative integration part is designed, which can adaptively aggregate the influence of multiple cues on dynamic saliency prediction.

1.3. Organization

The remaining sections of the paper are arranged as follows: In Section 2, the related works are reviewed. In Section 3, we describe the proposed method in details. In Section 4, the experimental results are shown and discussed. And in Section 5, the conclusion is drawn.
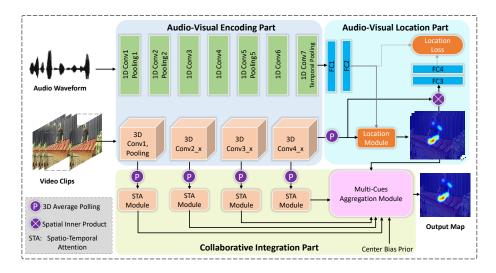
Figure 1: The framework of the proposed method. Firstly, an audio-visual encoding part is designed to learn audio-visual features. Secondly, an audio-visual location part is devised to locate the sound source in the visual scene by learning the correspondence between audio-visual information. Thirdly, a collaborative integration part is proposed to adaptively aggregate audio-visual information and center-bias prior to generate the final dynamic saliency map.

2. Related works

In recent years, numerous saliency prediction methods have been proposed. According to the data type, the existing methods can be divided into two aspects, including Static Saliency Prediction (SSP) methods and Dynamic Saliency Prediction (DSP) methods. This section reviews the existing saliency prediction methods in the following successively.

2.1. Static Saliency Prediction Methods

Earlier works devote to investigate the saliency prediction task based on static images [28–31]. These works are mostly on the basis of bottom-up visual attention mechanism [28, 32, 33]. Itti *et al.* [9] firstly conducted the saliency prediction task by imitating the human bottom-up visual selective attention process to extract the low-level visual features of images. To match actual eye movements, Judd *et al.* [33] considered both bottom-up visual attention and top-down image semantics, and collect a large eye tracking dataset to address the problem. On the assumption of contrast, many computational methods on SSP have been proposed. Perazzi *et al.* [34] designed an intuitive

SSP method based on the contrast feature. Wang *et al.* [35] leveraged selective contrast, including color, texture, and location, to predict the salient regions of images. In addition, many researchers have realized the SSP from the aspects of information-theoretic [10], decision-theoretic [11], and spectral analysis [12], *etc.*. Since the adopted features in these methods are hand-crafted, large-scale data with complex distributions cannot be well processed.

As a result, most researchers begin to conduct the SSP task by mining high-level semantic information in images with deep neural networks [13, 18, 19]. Vig *et al.* [36] firstly proposed utilizing deep neural networks for the SSP task. However, due to the insufficient training data, the performance is limited. To address this problem, Jiang *et al.* [37] built the SALICON dataset with plenty of natural images and the corresponding eye-tracking data. Based on the dataset, lots of works on SSP are developed by the follow-up researchers [38]. He *et al.* [38] explored the intrinsical reason of the large gap between deep models and the inter-human baseline. In addition, more effective network architectures are exploited for learning representative features [29–31, 39–41]. Zhang *et al.* [30] proposed to incorporate prior knowledge of semantic relationships so as to learn highlighted regions in images. Kroner *et al.* [31] developed an encoder-decoder structure to learn multi-scale high-level visual features for SSP. Considering that both low-level contrast features and high-level semantic features are important for SSP, Yuan *et al.* [42] introduced a bio-inspired representation learning method to generate the saliency map. Wang *et al.* [39] conducted the SSP by fusing features from multiple layers of VGG-16. Kruthiventi *et al.* [43] considered the centre-bias prior information and developed a computational method for SSP, which improves the predicted results considerably.

2.2. Dynamic Saliency Prediction Methods

With the tremendous progress of data storage technology and mobile Internet technology, massive video data are soaring recently [44, 45]. In order to deal with the vast amounts of information, researchers pay more attention on DSP to predict the most valuable information in video data [46–53]. Jiang *et al.* [46] developed an object-to-motion convolutional neural network for predicting the intra-frame visual saliency. Bak

6

et al. [47] presented a spatio-temporal saliency network for DSP. Gorji *et al.* [48] proposed a multi-stream ConvLSTM structure with the attentional push effect of the scene actors and the photographer. To improve the performance of DSP models, Sun *et al.* [49] put forward a step-gained fully convolutional network by simultaneously considering motion and temporal information. By conducting multi-scale feature learning and spatio-temporal feature integration, Lai *et al.* [50] designed a residual attentive learning network for DSP. Wu *et al.* [51] presented an end-to-end neural network named SalSAC for DSP. The network is based on CNN-LSTM-Attention and integrates both static and dynamic information. In order to fully consider the effect of both global and local consistency on DSP, Wang *et al.* [52] introduced a dynamic saliency network on the basis of both global discriminations and local consistency. These methods have greatly promoted the progress of DSP. Nevertheless, the effect of the audio information accompanying the video is ignored when conducting the DSP task.

Considering the influence of audio cues on the vision task, a few attempts have been made to better perceive the scene information and predict the saliency map. Starting from application-specific, some researcher adopt the traditional signal processing techniques for locating the salient region in the scene [54–56]. For example, Min *et al.* [56] utilized cross-modal kernel canonical correlation analysis to predict the moving-sounding object. Subsequently, more and more attention is paid on salient regions location by integration of audio [25, 57–61]. Qian *et al.* [25] proposed a two-stage audio-visual learning method for visually localizing multiple sound sources in unconstrained videos. To locate sounding objects in cocktail-party, Hu *et al.* [58] introduced a two-stage learning framework with a self-supervised class-aware manner. Afouras *et al.* [59] develop a model using attention to transform a video into several discrete audio-visual objects. Tavakoli *et al.* [62] designed a conceptually simple and effective audio-visual analysis method for dynamic saliency prediction. Tsiami *et al.* [60] proposed a spatio-temporal audio-visual saliency network by combining both visual and auditory information.

This paper is dedicated to conduct the dynamic saliency prediction task by exploring the collaborative mechanism among different audio-visual cues, while alleviating existing inconsistencies between audio-visual modalities.

7

3. The proposed network

In this work, an audio-visual collaborative representation learning method is proposed to generate dynamic saliency map. As is shown in Figure 1, the proposed method is composed of three main parts: 1) audio-visual encoding, 2) audio-visual location, and 3) collaborative integration parts. Thereinto, the audio-visual encoding part is responsible for learning audio-visual features, which contain spatial location and temporal motion information. The audio-visual location part aims at locating the sound source in the visual scene by learning the correspondence between audio-visual features. The collaborative integration part is in charge of adaptively aggregating the localized sound source information, spatio-temporal visual information (motion, high-level semantics, *etc.*), and center-bias prior information, so as to generate the final dynamic saliency map. These parts are expatiated successively as follows.

3.1. Audio-Visual Encoding

The audio-visual encoding part consists of two paralleled branches for encoding the original data as audio semantic features and spatio-temporal visual features, respectively. Details about the two branches are elaborated in the following subsections.

3.1.1. Audio Encoding

In the DSP task based on audio-visual analysis, it is important to obtain the semantic concept of audio rather than low-level signal [63]. To this end, the audio signals are represented by using convolutional neural networks (CNNs). Specifically, we follow the previous work [60] and adopt the 1-D fully convolutional network for processing the audio waveform. Firstly, the audio segment is cropped to match the visual frames duration (*i.e.* 16 frames). Secondly, a Hanning window is leveraged to acquire the central audio value with a higher weight, which represents the current time instance, and model the past and future attenuation values. Thirdly, a 1-D fully convolutional network with the first seven layers of the SoundNet [26] and a temporal max-pooling layer is applied for encoding high-level semantic features. Formulaically, the process

8

of audio encoding can be written as:

$$\mathbf{S}_A = \mathcal{F}_A(\mathbf{X}_A; \theta_A), \qquad (1)$$

where $\mathbf{X}_A$ and $\mathbf{S}_A$ represent the input audio data and the corresponding high-level semantic feature, respectively. $\mathcal{F}_A$ stands for the mapping function from audio data to the corresponding high-level semantic feature. θ_A is the parameter during the process of audio encoding.

3.1.2. *Spatio-Temporal Visual Encoding*

In order to capture the spatial semantic information and temporal motion information, the 3-D CNNs is adopted for processing the video frames. Specifically, the 3D ResNet-50 architecture [27], which is proposed initially for action recognition, is employed as the backbone to encode the spatio-temporal features. Lots of works [64–66] have demonstrated that multi-scale features contribute to achieving a good performance for perceiving objects with different scales. As a result, the multi-scale features are introduced for DSP in this work. As is shown in Figure 1, the spatio-temporal visual encoding branch adopts the first 4 ResNet convolutional blocks to provide the outputs $\mathbf{S}_V^1, \mathbf{S}_V^2, \mathbf{S}_V^3, \mathbf{S}_V^4$, which contain different spatial and temporal information. The process of spatio-temporal visual encoding can be written as:

$$\mathbf{S}_V^m = \mathcal{F}_V^m(\mathbf{X}_V; \theta_V^m), \qquad (2)$$

where $\mathbf{X}_V$ and $\mathbf{S}_V^m$ stand for the input video frames and the corresponding spatio-temporal feature of the m-th ResNet convolutional block. $\mathcal{F}_V^m$ is the mapping function from video frames to the corresponding spatio-temporal feature. θ_V^m is the parameter during the process of spatio-temporal visual encoding.

3.2. *Audio-Visual Location*

In this subsection, an audio-visual location part is devised to locate the sounding object by exploiting the consistency between the audio and visual modalities in a sharing latent space. By this way, the related sounding objects are selectively and dynamically

brought out to the foreground when audio and visual concepts appear simultaneously. For example, the playing the piano, barking dog, *etc.*, in the video, are expected to be detected after the audio-visual location.

The audio-visual location can be implemented with 4 steps. Firstly, the output of the 4-th ResNet convolutional block is selected as spatio-temporal visual feature, termed as $\mathbf{S}_V^4$, since it contains rich semantic information for the visual frames. In order to marginal out the temporal dimension and acquire a global representation, a temporal average pooling operation is applied for $\mathbf{S}_V^4$. For simplicity, the global representation is denoted as a reshaped matrix form $\mathbf{V} = [\mathbf{v}_1; \cdots; \mathbf{v}_B] \in \mathbb{R}^{B \times D_h}$. Secondly, to match the dimension of the global representation in visual branch and higher level concept of audio signal, two fully connected (FC) layers with ReLU activation are applied for the audio feature $\mathbf{S}_A$, so as to generate the audio embedding $\mathbf{h}_A$ with D_h-dimension. Thirdly, an attention mechanism is utilized in the location module (see Figure 1) to locate the sounding object and generate a audio-aware saliency map. It is achieved as follows:

$$a_b = \langle \mathbf{W}_1 \mathbf{v}_b, \mathbf{W}_2 \mathbf{h}_A \rangle, \tag{3}$$

$$\alpha_b = \frac{\exp(a_b)}{\sum_{b=1}^{B} \exp(a_b)}, \tag{4}$$

where a_b captures the dependency between $\mathbf{h}_A$ and $\mathbf{v}_b$. $\mathbf{W}_1$ and $\mathbf{W}_2$ are the training weights. $\langle \cdot, \cdot \rangle$ means the inner-product operation between two matrixes. α_b stands for the b-th element in the sounding map $\boldsymbol{\alpha}$, which can be interpreted as the probability of location related to the audio context. Further, the audio-aware saliency map $\mathbf{F}_{\text{audio}}$ can be computed by a upsampling operation. Note that the upsampling operation adopts the resize-convolution operation rather than deconvolution to avert the checkerboard effect[1]. After obtaining the sounding map, we fourthly compute the representative context vector $\mathbf{h}_z$ to interact the sounding map with the global representation in visual

_____

[1] https://distill.pub/2016/deconv-checkerboard/

10

branch at the sound source location. The process is achieved by:

$$\mathbf{h}_z = \sum_{b=1}^{B} \alpha_b \mathbf{v}_b. \tag{5}$$

Then, the representative context vector $\mathbf{h}_z$ is transformed to a spatio-temporal visual feature $\hat{\mathbf{v}}$ with two fully connected (FC) layers with ReLU activation. Finally, a location loss is imposed on $\hat{\mathbf{v}}$ and $\mathbf{S}_A$ to learn the features to share latent space of the audio and visual modality. Here, the Euclidean distance is employed to calculate the location loss.

3.3. Collaborative Integration

As mentioned before, human attention is influenced by many aspects, including the sounding, moving, center, high-level semantic objects, *etc.*. To model these aspects, this subsection presents a collaborative integration mechanism. Specifically, a spatio-temporal attention module is proposed to detect the moving and semantic objects. A learnable center-bias prior function is introduced to generate center-bias prior-aware map. A multi-cues aggregation module is devised to integrate the influence of different cues and generate the final saliency map.

3.3.1. *Moving and High-Level Semantic Objects Prediction*

In order to model the influence of moving and semantic objects for DSP, a spatio-temporal attention module is proposed. The spatio-temporal attention module contains two branches for capturing the moving information and semantic information, respectively.

As for the temporal attention branch, the temporally moving information is modeled. As is shown in Figure 2 (a), the spatio-temporal visual feature from m-th ResNet block is firstly processed with the operation of averaging pooling in the channel dimension, resulting in $\mathbf{S}_{V,T}^m$. Secondly, the first and last frames are removed, respectively, leading to $\mathbf{S}_{V,T,-1}^m$ and $\mathbf{S}_{V,T,0}^m$. Thirdly, the temporal attention is computed by conducting the frame-wise similarity between $\mathbf{S}_{V,T,-1}^m$ and $\mathbf{S}_{V,T,0}^m$, so as to capture the temporally moving information. Specifically, the frame-wise similarity can be imple-
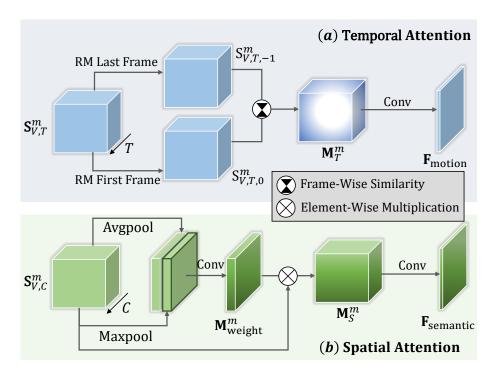
Figure 2: The devised spatio-temporal attention module.

mented as:

$$\mathbf{M}_T^m = \sum_{t=1}^{t=T-1} \left(1 - \left(\mathbf{S}_{V,t,0}^m - \mathbf{S}_{V,t,-1}^m\right)\right), \tag{6}$$

where $\mathbf{M}_T^m$ represent the motion feature from the m-th ResNet convolutional block. t indexes the frames. Afterwards, the motion feature is further processed with a 1×1 convolution and a resize-convolution operation to generate a motion-aware saliency map $\mathbf{F}_{\text{motion}}^m$.

In respect to the spatial attention branch, the high-level semantic information is expected to be captured by leveraging the inter-spatial relationship of features. As is shown in Figure 2 (b), the spatio-temporal visual feature from m-th ResNet block is firstly processed with the operation of averaging pooling in the temporal dimension, so as to obtain $\mathbf{S}_{V,C}^m$. Secondly, we apply the max-pooling and average-pooling along with the channel dimension, and concatenate them as an efficient feature descriptor. In this way, the highlighting informative regions can be effectively shown [67, 68].

Thirdly, a convolution layer with filter size of 7×7 is conducted for generating a spatial weight matrix $\mathbf{M}^m_{\text{weight}}$. Fourthly, by employing an element-wise multiplication between $\mathbf{S}^m_{V,C}$ and the spatial weight matrix $\mathbf{M}^m_{\text{weight}}$, the feature $\mathbf{M}^m_S$ with high-level semantic information is learned. Finally, $\mathbf{M}^m_S$ is further processed with the 1×1 convolution and resize-convolution operation to generate a high-level semantic-aware saliency map $\mathbf{F}^m_{\text{semantic}}$.

3.3.2. *Center-Bias Prior*

According to previous studies [30, 69, 70], human attention tends to concentrate on the center of scenes, which is termed as center-bias phenomenon. To this end, a learnable center-bias prior function is adopted according to our preceding work [42]. Specifically, the center-bias prior-aware map $\mathbf{F}_{\text{center}}$ is generated using a Gaussian function as follows:

$$\mathbf{F}_{\text{center}} = \frac{1}{2\pi\sigma_x\sigma_y}\exp\left(-\left(\frac{(x-x_0)^2}{2\sigma_x^2} + \frac{(y-y_0)^2}{2\sigma_y^2}\right)\right), \quad (7)$$

where σ_x^2 and σ_y^2 indicate the to-be-learned horizontal variance and vertical variance, respectively. The generated center-bias prior-aware map $\mathbf{F}_{\text{center}}$ represents a spatial pattern. Note that the center-bias information is modeled purely from learning.

3.3.3. *Multi-Cues Aggregation*

We have computed the audio-aware saliency map, the motion-aware saliency map, the high-level semantic-aware saliency map, and the center-bias prior-aware saliency map. They can express the saliency driven by different cues. As a result, it is quite essential to integrate them for generating the final saliency map. For this purpose, a multi-cues aggregation module is proposed for integrating the influence of different cues, by exploiting the consistency among them and reducing the difference.

As is shown in Figure 3, the multi-cues aggregation is conducted by two branches. One branch is to learn the global channel context, and the other branch is responsible for perceiving local channel context. The outputs of the two branches are combined to obtain the fused multi-cues feature. Specifically, the multi-cues aggregation

13

module is composed of three main steps. Firstly, the concatenated feature $\mathbf{M}_{\mathrm{conc}} = [\mathbf{F}_{\mathrm{audio}}; \mathbf{F}_{\mathrm{motion}}; \mathbf{F}_{\mathrm{semantic}}; \mathbf{F}_{\mathrm{center}}]$ is processed as the global response context $\mathbf{g}$ by:

$$
\begin{aligned}
\mathbf{g} &= \mathrm{global}(\mathbf{M}_{\mathrm{conc}}; \mathbf{W}_3) \\
&= \sigma\left(\mathcal{B}\left(\mathrm{PWC}\left(\sigma\left(\mathcal{B}\left(\mathrm{PWC}\left(\mathrm{GAP}(\mathbf{M}_{\mathrm{conc}})\right)\right)\right)\right)\right)\right),
\end{aligned}
\tag{8}
$$

where global denotes the global response mapping function. $\mathbf{W}_3$ is the to-be-learned parameter. σ represents the Sigmoid function. $\mathcal{B}$ means the Batch Normolization (BN) operation. PWC stands for the Point-Wise Convolution (PWC), which is chosen for its lightweight. GAP represents the global average pooing.

In parallel, the local response context $\mathbf{L}$ can be computed by:

$$
\begin{aligned}
\mathbf{L} &= \mathrm{local}(\mathbf{M}_{\mathrm{conc}}; \mathbf{W}_4) \\
&= \mathcal{B}\left(\mathrm{PWC}\left(\delta\left(\mathcal{B}\left(\mathrm{PWC}\left(\mathbf{M}_{\mathrm{conc}}\right)\right)\right)\right)\right),
\end{aligned}
\tag{9}
$$

where local denotes the local response mapping function. $\mathbf{W}_4$ is the to-be-learned parameter. δ represents the ReLU function.

Secondly, to better integrate the influence of different cues on DSP, the global response context and local response context are combined as:

$$
\mathbf{M}_{\mathrm{fusion}} = \mathbf{g} \odot \mathbf{L},
\tag{10}
$$

where $\mathbf{M}_{\mathrm{fusion}}$ represents the fused feature by considering different cues, and $\odot$ means the channel-wise multiplication.

Finally, based on the fused feature $\mathbf{M}_{\mathrm{fusion}}$, the final saliency map $\mathbf{F}_{\mathrm{map}}$ is computed by a Readout Network, which is composed of three successive 1×1 convolution layers.

3.4. Optimizing Strategy

In order to obtain the final saliency map, we aggregate the saliency map driven by cues of sounding, moving, center, high-level semantic objects. Let $\mathbf{Y}_{sal}$ represent
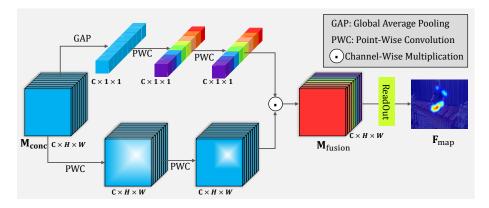
Figure 3: The proposed multi-cues aggregation module.

the ground-truth fixation map obtained by the eye-tracking data. During audio-visual location process, the objective function is defined as:

$$
\begin{aligned}
\mathcal{L}_{\mathrm{A}} &= \mathrm{KL}\left(\mathbf{F}_{\mathrm{audio}}, \mathbf{Y}_{\mathrm{sal}}\right) + \|(\hat{\mathbf{v}}, \mathbf{S}^{A})\|_{2} \\
&= \mathbf{Y}_{\mathrm{sal}} \log\left(\frac{\mathbf{Y}_{\mathrm{sal}}}{\mathbf{F}_{\mathrm{audio}} + \epsilon} + \epsilon\right) + \|(\hat{\mathbf{v}}, \mathbf{S}^{A})\|_{2},
\end{aligned}
\tag{11}
$$

where $\mathrm{KL}(\cdot, \cdot)$ stands for the Kullback–Leibler divergence between the two distributions. $\|\cdot, \cdot\|_{2}$ is the Euclidean distance. ϵ indicates the regularization constant to avoid the NaN value in the loss.

During the moving and high-level semantic objects prediction process, the objective function is defined as:

$$
\begin{aligned}
\mathcal{L}_{\mathrm{MS}} =& \mathrm{KL}\left(\mathbf{F}_{\mathrm{motion}}, \mathbf{Y}_{\mathrm{sal}}\right) + \mathrm{KL}\left(\mathbf{F}_{\mathrm{semantic}}, \mathbf{Y}_{\mathrm{sal}}\right) \\
=& \mathbf{Y}_{\mathrm{sal}} \log\left(\frac{\mathbf{Y}_{\mathrm{sal}}}{\mathbf{F}_{\mathrm{motion}} + \epsilon} + \epsilon\right) \\
& + \mathbf{Y}_{\mathrm{sal}} \log\left(\frac{\mathbf{Y}_{\mathrm{sal}}}{\mathbf{F}_{\mathrm{semantic}} + \epsilon} + \epsilon\right).
\end{aligned}
\tag{12}
$$

During the multi-cues aggregation process, the objective function is defined as:

$$
\begin{aligned}
\mathcal{L}_{\mathrm{fuse}} &= \mathrm{KL}\left(\mathbf{F}_{\mathrm{map}}, \mathbf{Y}_{\mathrm{sal}}\right) \\
&= \mathbf{Y}_{\mathrm{sal}} \log\left(\frac{\mathbf{Y}_{\mathrm{sal}}}{\mathbf{F}_{\mathrm{map}} + \epsilon} + \epsilon\right).
\end{aligned}
\tag{13}
$$

15

Ultimately, the final loss $\mathcal{L}_{\text{final}}$ of training all parameters can be jointly combined by the losses $\mathcal{L}_A$, $\mathcal{L}_{MS}$, and $\mathcal{L}_{fuse}$, as follows.

$$\mathcal{L}_{\text{final}} = w_1\mathcal{L}_{\text{A}} + w_2\mathcal{L}_{\text{MS}} + w_3\mathcal{L}_{\text{fuse}}, \tag{14}$$

where w_1, w_2, and w_3 are the tradeoff coefficients controlling the contribution of each term.

Algorithm 1 The proposed method

Input:
 Training video frames $\mathbf{X}_V$ and the corresponding audio data $\mathbf{X}_A$;
 Testing video frames $\mathbf{X}_V^{te}$ and the corresponding audio data $\mathbf{X}_A^{te}$.

Output:
 Testing saliency map $\mathbf{F}_{\text{map}}^{te}$;
 All the to-be-learned parameters $\mathbf{W}$.

Initialization:
 The parameter θ_A in the audio encoding branch is initialized from the origin SoundNet [26]. The parameter θ_V in the spatio-temporal visual encoding branch is initialized from the origin 3D ResNet-50 [27]. The remaining weights are randomly initialized by truncated_normal distribution.

Repeat:
 1: Calculate the high-level semantic feature $\mathbf{S}_A$ and spatio-temporal feature $\mathbf{S}_V$ according to Eq. 1 and Eq. 2, respectively;
 2: Generate the audio-aware saliency map $\mathbf{F}_{\text{audio}}$ according to Eq. 3 and Eq. 4;
 3: Calculate the motion-aware saliency map $\mathbf{F}_{\text{motion}}$, the high-level semantic-aware saliency map $\mathbf{F}_{\text{semantic}}$, and center-bias prior-aware map $\mathbf{F}_{\text{center}}$ based on Section 3.3.1 and Section 3.3.2;
 4: Generate the final saliency map $\mathbf{F}_{\text{map}}$ based on Section 3.3.3;
 5: Compute the final loss $\mathcal{L}_{\text{final}}$ according to Eq. 14;
 6: Update all the parameters by utilizing Adam optimizer.

Until: A fixed number of iterations.
 7: Generate the testing saliency map $\mathbf{F}_{\text{map}}^{te}$.

Return: $\mathbf{F}_{\text{map}}^{te}$, $\mathbf{W}$

Based on the final loss $\mathcal{L}_{\text{final}}$, the proposed method can be optimized as follows. The parameter θ_A in the audio encoding part are initialized from the origin SoundNet [26]. The parameter θ_V in the spatio-temporal visual encoding part is initialized from the origin 3D ResNet-50 [27], which is pretrained on the Kinetics 400 dataset for action recognition task. The remaining weights are randomly initialized by truncated_normal distribution. In the training stage, the optimizing process is composed of five main

steps. Firstly, the training video frames $\mathbf{X}_V$ and the corresponding audio data $\mathbf{X}_A$ are processed as high-level semantic feature $\mathbf{S}_A$ and spatio-temporal feature $\mathbf{S}_V$ with the audio-visual encoding part. Secondly, the high-level semantic feature $\mathbf{S}_A$ and spatio-temporal feature $\mathbf{S}_V$ are combined to locate the sounding object to generate the audio-aware saliency map $\mathbf{F}_{\mathrm{audio}}$ by exploiting the consistency in a sharing latent space of the audio and visual modality. Thirdly, the motion-aware saliency map $\mathbf{F}_{\mathrm{motion}}$, the high-level semantic-aware saliency map $\mathbf{F}_{\mathrm{semantic}}$, and center-bias prior-aware map $\mathbf{F}_{\mathrm{center}}$ are computed based on Section 3.3.1 and Section 3.3.2; Fourthly, the final saliency map $\mathbf{F}_{\mathrm{map}}$ are inferred based on $\mathbf{F}_{\mathrm{audio}}$, $\mathbf{F}_{\mathrm{motion}}$, $\mathbf{F}_{\mathrm{semantic}}$, and $\mathbf{F}_{\mathrm{center}}$. Finally, we compute the final loss $\mathcal{L}_{\mathrm{final}}$ based on the generated final saliency map $\mathbf{F}_{\mathrm{map}}$ and the ground-truth fixation map $\mathbf{Y}_{\mathrm{sal}}$, and update all the parameters $\mathbf{W}$ by minimizing $\mathcal{L}_{\mathrm{final}}$ with Adam optimizer. Once the training epoch reaches 50, the training process is terminated. Afterwards, the parameter $\mathbf{W}$ is utilized to infer the testing saliency map $\mathbf{F}_{\mathrm{map}}^{te}$. It is to note that the proposed method is trained in an end-to-end manner. The details about the optimization process are shown in Algorithm 1.

4. Experiment and results

The experiments are conducted on six benchmark datasets with audio-visual eye-tracking data. In the following subsections, the implementation details, evaluation metrics are elaborated. In addition, the experimental results are given and analyzed from the aspects of ablation study and comparison with the state-of-the-arts.

4.1. Setup

4.1.1. Datasets

The proposed method is trained and evaluated on AVAD [54], Coutrot1 [71, 72], Coutrot2 [71, 72], DIEM [73], ETMD [74, 75] and SumMe [75, 76] datasets. These datasets contains various types videos accompanied with audios. Specifically, 1) the AVAD dataset [54] contains 45 video clips with each duration 5-10 seconds. It covers various audio-visual activities, *e.g.*, playing the piano, playing basketball, making an interview, *etc.*. The dataset also contains the eye-tracking data from 16 participants.

2) The Coutrot1 and Coutrot2 datasets are split from the Coutrot dataset [71, 72]. The Coutrot1 dataset is with 60 video clips covering 4 visual categories: one moving object, several moving objects, landscapes, and faces. The corresponding eye-tracking data are recorded from 72 participants. The Coutrot2 dataset includes 15 video clips, which record 4 persons having a meeting. The corresponding eye-tracking data are from 40 participants. 3) The DIEM dataset [73] involves 84 video clips, including game trailers, music videos, advertisements, *etc.*. Note that the audio and visual tracks do not correspond naturally. The eye-tracking data are recorded via 42 participants. The ETMD dataset [74, 75] includes 12 video clips from several hollywood movies. The eye-tracking data are annotated by 10 different people. The SumMe dataset [75, 76] consists of 25 video clips with diverse topics, *e.g.*, playing ball, cooking, traveling, *etc.*. The corresponding eye-tracking data are collected from 10 viewers.

Following the previous work [60], we adopt the same data partitioning for traning and testing. Specifically, 3 different splits of the data are created with non-overlapping among train, validation and test sets. The performance are evaluated by taking the average among all 3 splits.

4.1.2. *Evaluation Metrics*

In order to measure the consistency between the predicted saliency map and the groundtruth fixation map, 5 widely-used evaluation metrics for DSP are employed [77]. The evaluation metrics include CC, NSS, AUC-Judd (AUC-J), shuffled AUC (sAUC), and SIM. The CC measures the linear correlation coefficient between the groundtruth fixation map and the predicted saliency map. The NSS aims at measuring the saliency value on human fixations. The AUC-J and sAUC are location-based metrics for evaluate the predicted saliency map. The SIM measures the similarity between the predicted saliency map and groundtruth fixation map. The 5 evaluation metrics provide a comprehensive assessment for DSP.

4.1.3. *Implementation Details*

The input samples are processed as 16 video frames and the corresponding audio stream. Each video frame is resized at 112×112 pixels. Following the previous work

18

[60], the data augmentation is also employed for random generation of training samples. The implementation adopts the 3D ResNet-50 [27] as backbone for encoding spatio-temporal visual features, and applies SoundNet [26] as backbone for encoding high-level audio semantic features. The Gaussian kernel for generating center-bias prior map is with size of 7×7. The tradeoff coefficients w_1, w_2, and w_3 in Eq. 14 are all selected as 1 empirically. The batchsize is set as 128. The proposed method is optimized by utilizing the Adam optimizer with learning rate of 10^{-4}. When the iterative epoch reaches 50, the optimizing process is terminated. During the test process, a sliding window method is adopted for inferring the final saliency map of each frame. The experiment is conducted by the $Pytorch$ library and on the PC with a TITAN RTX GPU and 24G RAM.

4.2. Ablation Analysis

In this subsection, we aims at verifying the effectiveness of several main parts for the proposed method. Specifically, six different variations are constructed, including:

- **Visual Model:** Only visual information is leveraged for DSP, while the audio information is ignored.

- **AV Inner-Product:** The audio-visual location is implemented by directly conducting inner-product operation, rather than by adopting the proposed audio-visual location part to exploit the consistency between the audio and visual modalities in a sharing latent space.

- **Concatenate Fusion:** In order to integrate the multi-cues maps and generate the final saliency map, we directly concatenate them and further readout, rather than by utilizing the proposed multi-cues aggregation module.

- **Proposed (w/o SA):** The spatial attention is not considered for modeling higher-level semantic information.

- **Proposed (w/o TA):** The temporal attention is not considered for modeling motion information.

- **Proposed:** The complete method proposed by us.

Table 1: The ablation study on six benchmark datasets.

Methods	AVAD					Coutrot1				
	CC↑	NSS↑	AUC-J↑	sAUC↑	SIM↑	CC↑	NSS↑	AUC-J↑	sAUC↑	SIM↑
Visual Model	0.5714	3.11	0.8928	0.5740	0.4635	0.4458	2.03	0.8527	0.5562	0.3630
AV Inner-Product	0.5751	3.14	0.9004	0.5827	0.4572	0.4608	2.16	0.8604	0.5681	0.3744
Concatenate Fusion	0.5826	3.32	0.9081	0.5944	0.4680	0.4714	2.29	0.8702	0.5774	0.3853
Proposed (w/o SA)	0.5907	3.36	0.9127	0.6014	0.4752	0.4807	2.35	0.8696	0.5842	0.3967
Proposed (w/o TA)	0.6125	3.44	0.9204	0.6145	0.4783	0.4852	2.36	0.8704	0.5861	0.4017
Proposed	**0.6262**	**3.57**	**0.9251**	**0.6203**	**0.4820**	**0.4985**	**2.44**	**0.8798**	**0.6042**	**0.4154**

Methods	Coutrot2					DIEM				
	CC↑	NSS↑	AUC-J↑	sAUC↑	SIM↑	CC↑	NSS↑	AUC-J↑	sAUC↑	SIM↑
Visual Model	0.4260	3.26	0.9208	0.6344	0.3071	0.5461	2.08	0.8544	0.6351	0.4335
AV Inner-Product	0.6748	4.68	0.9370	0.6727	0.4283	0.5542	2.15	0.8716	0.6552	0.4551
Concatenate Fusion	0.7034	4.91	0.9435	0.6911	0.4824	0.5763	2.21	0.8835	0.6740	0.4673
Proposed (w/o SA)	0.7221	5.09	0.9466	0.7028	0.4960	0.5816	2.25	0.8847	0.6781	0.4726
Proposed (w/o TA)	0.7283	5.14	0.9473	0.7135	0.5048	0.5834	2.29	0.8903	0.6844	0.4830
Proposed	**0.7481**	**5.45**	**0.9537**	**0.7294**	**0.5266**	**0.5924**	**2.33**	**0.8941**	**0.6982**	**0.4917**

Methods	ETMD					SumMe				
	CC↑	NSS↑	AUC-J↑	sAUC↑	SIM↑	CC↑	NSS↑	AUC-J↑	sAUC↑	SIM↑
Visual Model	0.4817	2.41	0.9180	0.6842	0.3504	0.3482	1.86	0.8573	0.6235	0.2840
AV Inner-Product	0.5024	2.53	0.9247	0.6981	0.3584	0.3615	1.92	0.8650	0.6361	0.3015
Concatenate Fusion	0.5346	2.81	0.9283	0.7217	0.4011	0.4037	2.16	0.8742	0.6527	0.3281
Proposed (w/o SA)	0.5284	2.73	0.9217	0.7185	0.3952	0.4164	2.18	0.8780	0.6613	0.3327
Proposed (w/o TA)	0.5415	2.99	0.9274	0.7261	0.4133	0.4228	2.21	0.8807	0.6653	0.3371
Proposed	**0.5664**	**3.05**	**0.9351**	**0.7406**	**0.4325**	**0.4392**	**2.25**	**0.8945**	**0.6712**	**0.3428**

TABLE 1 exhibits the results of different variations. By the observation and analysis from the results, we can verify five main observations:

1) The audio information plays a significant role on DSP. The conclusion can be supported by the comparison results between Visual Model and the proposed method. Specifically, It is worth noticing that the performance drops significantly when only visual information is leveraged. Concretely, the CC value drops by more than 5% on the AVAD dataset. The NSS value is dropped form 5.45 to 3.26 on the Coutrot2 dataset. The AUC-J value is decreased from 0.8941 to 0.8544 on the DIEM dataset. The sAUC value falls almost 6% on the ETMD dataset. And the SIM metric drops nearly 8% on the SumMe dataset. The results demonstrate the important role of audio information for the DSP task. Especially, the results on the Coutrot2 dataset are more able to illustrate this point, since the Coutrot2 records 4 persons having a meeting, and the audio plays a great role on human attention. In addition, Figure 4 depicts some visualized results. Expectedly, from the comparison results of the third row and the fourth row, we can clearly observe that the predicted saliency map is more accurate when the audio information is considered. Concretely, in the first video (the first two columns), when the audio information is ignored, the visual model locates the non-sounding person,

Figure 4: Some examples by adopting different settings. The first row shows the raw frames. The second row shows the corresponding groundtruth (GT) of the saliency map. The third row exhibits the predicted saliency maps when audio and visual information is employed simultaneously. The last row displays the predicted saliency maps when only visual information is leveraged.

which is no salient. This further demonstrate that the audio information can prompt the model to better locate the attention-grabbing sounding objects in the scene.

2) The devised audio-visual location part can locate the sounding objects effectively so as to enhance the performance of DSP methods. For the observation from TABLE 1, we can find that when the devised audio-visual location is replaced with inner-product operation, the performance for DSP declines notably. For instance, the CC value is decreased from 0.6262 to 0.5151 on the AVAD dataset, when we adopt the AV Inner-Product method. This is because the devised audio-visual location part can be utilized to effectively locate the sounding objects, which more attract human attention.

3) The proposed multi-cues aggregation module can be capable to integrate the effects by multiple cues. From TABLE 1, we can find the great differences between the results from Concatenate Fusion method and the proposed method. It is mainly because the Concatenate Fusion method can not effectively integrate the influences of multiple cues on saliency. In contrast, the proposed method can achieve this effectually, which indicates the proposed multi-cues aggregation module is able to integrate the effects by multiple cues.

4) The spatial attention contributes to model the higher-level semantic information for better DSP. Based on the comparison results between Proposed (w/o SA) and the proposed method, we can discover that the latter can get more superior performance on all benchmark datasets. Especially, the CC value on the ETMD dataset can be improved nearly 4%, which indicates the importance of the spatial attention for capturing the higher-level semantic information.

5) The temporal attention lifts the performance of DSP methods by modeling motion information. The comparison results between Proposed (w/o TA) and the proposed method on TABLE 1 reveals the significance of the adopted temporal attention. It is because the temporal attention can capture the temporally moving information by exploiting the frame-wise similarity. The results verify the importance of the temporal attention for improving the performance of DSP methods.

4.3. Comparison with State-of-the-arts

To demonstrate the effectiveness of the proposed method, we compare the proposed method with 8 state-of-the-art DSP methods. These comparison methods are comprehensive, including 4 spatial methods and 5 spatio-temporal methods. 1) The spatial methods process each frame independently to generate the saliency map, while do not consider the temporal information among frames. The comparison spatial methods include DeepNet [78], DVA [13], SAM [69], and SalGAN [79] methods. DeepNet [78] addresses the DSP problem by utilizing convnets for regression in an end-to-end manner. DVA [13] learns multi-scale features for capturing hierarchical saliency information for DSP. SAM [69] predicts the saliency map by incorporating neural attentive mechanisms using a convolutional long short-term memory. SalGAN [79] adopts the data-driven metric for DSP by training with an adversarial loss function. 2) The spatio-temporal methods generate the saliency map by capturing the spatial information within each frame and the temporal information among frames simultaneously. The comparison spatial methods include ACLNet [80], DeepVS [46], TASED [81], and STAViS [60] methods. ACLNet [80] employs the CNN-LSTM architecture for DSP with a supervised attention mechanism. DeepVS [46] develops an object-to-motion convolutional neural network for estimating the intra-frame saliency. TASED

22

Table 2: The quantitative comparison on six benchmark datasets.

Methods	AVAD					Coutrot1				
	CC↑	NSS↑	AUC-J↑	sAUC↑	SIM↑	CC↑	NSS↑	AUC-J↑	sAUC↑	SIM↑
DeepNet [78]	0.3831	1.85	0.8690	0.5616	0.2564	0.3402	1.41	0.8248	0.5597	0.2732
DVA [13]	0.5247	3.00	0.8887	0.5820	0.3633	0.4306	2.07	0.8531	0.5783	0.3324
SAM [69]	0.5279	2.99	0.9025	0.5777	0.4244	0.4329	2.11	0.8571	0.5768	0.3672
SalGAN [79]	0.4912	2.55	0.8865	0.5799	0.3608	0.4161	1.85	0.8536	0.5799	0.3321
ACLNet [80]	0.5809	3.17	0.9053	0.5600	0.4463	0.4253	1.92	0.8502	0.5429	0.3612
DeepVS [46]	0.5281	3.01	0.8968	0.5858	0.3914	0.3595	1.77	0.8306	0.5617	0.3174
TASED [81]	0.6006	3.16	0.9146	0.5898	0.4395	0.4799	2.18	0.8676	0.5808	0.3884
STAViS [60]	0.6086	3.18	0.9196	0.5936	0.4578	0.4722	2.11	0.8686	0.5847	0.3935
Proposed	**0.6262**	**3.57**	**0.9251**	**0.6203**	**0.4820**	**0.4985**	**2.44**	**0.8798**	**0.6042**	**0.4154**

Methods	Coutrot2					DIEM				
	CC↑	NSS↑	AUC-J↑	sAUC↑	SIM↑	CC↑	NSS↑	AUC-J↑	sAUC↑	SIM↑
DeepNet [78]	0.3012	1.82	0.8966	0.6000	0.2019	0.4075	1.52	0.8321	0.6227	0.3183
DVA [13]	0.4634	3.45	0.9328	0.6324	0.2742	0.4779	1.97	0.8547	0.6410	0.3785
SAM [69]	0.4194	3.02	0.9320	0.6152	0.3041	0.4930	2.05	0.8592	0.6446	0.4261
SalGAN [79]	0.4398	2.96	0.9331	0.6183	0.2909	0.4868	1.89	0.8570	0.6609	0.3931
ACLNet [80]	0.4485	3.16	0.9267	0.5943	0.3229	0.5229	2.02	0.8690	0.6221	0.4279
DeepVS [46]	0.4494	3.79	0.9255	0.6469	0.2590	0.4523	1.86	0.8406	0.6256	0.3923
TASED [81]	0.4375	3.17	0.9216	0.6118	0.3142	0.5579	2.16	0.8812	0.6579	0.4615
STAViS [60]	0.7349	5.28	**0.9581**	0.7106	0.5111	0.5795	2.26	0.8838	0.6741	0.4824
Proposed	**0.7481**	**5.45**	0.9537	**0.7294**	**0.5266**	**0.5924**	**2.33**	**0.8941**	**0.6982**	**0.4917**

Methods	ETMD					SumMe				
	CC↑	NSS↑	AUC-J↑	sAUC↑	SIM↑	CC↑	NSS↑	AUC-J↑	sAUC↑	SIM↑
DeepNet [78]	0.3879	1.90	0.8897	0.6992	0.2253	0.3320	1.55	0.8488	0.6451	0.2274
DVA [13]	0.4965	2.72	0.9039	0.7288	0.3165	0.3983	2.14	0.8681	0.6686	0.2811
SAM [69]	0.5068	2.78	0.9073	0.7310	0.3790	0.4041	2.21	0.8717	0.6728	0.3272
SalGAN [79]	0.4765	2.46	0.9035	**0.7463**	0.3117	0.3978	1.97	0.8754	**0.6882**	0.2897
ACLNet [80]	0.4771	2.36	0.9152	0.6752	0.3290	0.3795	1.79	0.8687	0.6092	0.2965
DeepVS [46]	0.4616	2.48	0.9041	0.6861	0.3495	0.3172	1.62	0.8422	0.6120	0.2622
TASED [81]	0.5093	2.63	0.9164	0.7117	0.3660	0.4288	2.10	0.8840	0.6570	0.3337
STAViS [60]	**0.5690**	2.94	0.9316	0.7317	0.4251	0.4220	2.04	0.8883	0.6562	0.3373
Proposed	0.5664	**3.05**	**0.9351**	0.7406	**0.4325**	**0.4392**	**2.25**	**0.8945**	0.6712	**0.3428**

[81] exploits 3D fully-convolutional network architecture to generate the saliency map of each frame by considering several past frames. STAViS [60] combines both visual and auditory information for DSP in videos.

On the on hand, we compare the proposed method with 8 state-of-the-art DSP methods qualitatively on six benchmark datasets with audio-visual eye-tracking data, including AVAD [54], Coutrot1 [71, 72], Coutrot2 [71, 72], DIEM [73], ETMD [74, 75] and SumMe [75, 76] datasets. TABLE 2 reports the qualitative results. As can be observed evidently, the proposed method outperforms the comparison methods with respect to most evaluation metrics. Especially, the proposed method surpasses the spatial DSP methods, such as DeepNet [78], DVA [13], SAM [69], and SalGAN [79], by a substantial margin. The success attributes to the proposed method capturing the temporal information, which is quite important for DSP. Compared with the spatio-temporal DSP methods, *e.g.*, ACLNet [80], DeepVS [46], TASED [81], and STAViS
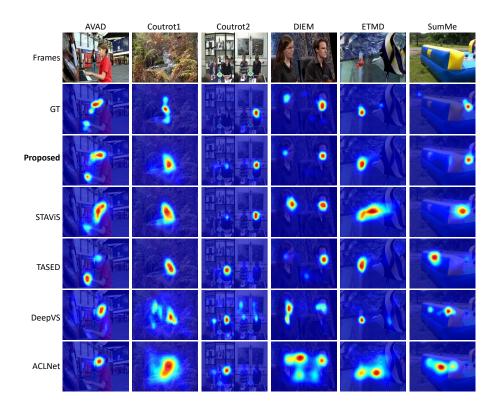
Figure 5: Qualitative results of different methods on six benchmark datasets. From top to bottom, each row represents the video frames, the corresponding groundtruth (GT), the predicted saliency map with the proposed method and other comparison methods.

[60], the proposed method also exhibits better performance. It is mainly because the proposed method can adaptively integrate multiple cues, which are essential for DSP. More specifically, even though the STAViS method also adopts audio-visual information for DSP, the proposed method surpasses it on the prediction of saliency map. This because we integrate more information affecting DSP, design more effective audio-visual location part for locating the sounding objects, and propose multi-cues aggregation module to collaboratively integrate the influence of multiple cues on dynamic saliency prediction. In addition, the higher CC and SIM values show that the generated saliency maps by the proposed method are more similar to the human annotations, which further demonstrates the superiority of the proposed method.

On the other hand, the qualitative visual comparisons are also conducted, and the

results are shown in Figure 5. As can be seen, the proposed method achieves the optimal performance among the comparison methods. Concretely, as for the sample in first column, the proposed method can locate both the sounding object (the piano) and high-level semantic object (human face) at the same time. In contrast, the TASED method mainly focuses on the sounding object. The DeepVS and ACLNet methods only pay attention to the high-level semantic object. The reason is that the proposed method collaboratively integrates multiple cues for DSP. Compared with the visualized results of STAVis, the proposed method can predict the attention-grabbing objects in the scene more accurately. It indicates the importance of other cues except for audio cues for DSP, and the superiority of audio-visual location part for locating sounding objects and the multi-cues aggregation module for adaptively integrating multi-cues information.

5. Conclusions

In this paper, we propose an audio-visual collaborative representation learning method for dynamic saliency prediction based on the fact that human attention is naturally influenced by audio-visual stimuli, including sounding, moving, center, high-level semantic objects. Specifically, the proposed method consists of three main parts: audio-visual encoding, audio-visual location, and collaborative integration parts. The audio-visual encoding part encodes the input data as audio semantic features and spatio-temporal visual features. The audio-visual location part locates the sounding object by exploiting the consistency between the audio and visual modalities in a sharing latent space. The collaborative integration part captures the moving, center, and high-level semantic information, and adaptively integrates them with the sounding information to generate the final saliency map. The experimental results on six benchmark datasets have demonstrated that: 1) the audio information contributes to the DSP task significantly; 2) the consistency between the audio and visual modalities in a sharing latent space enhances the performance of sounding object location; 3) the adaptively aggregation of multi-cues helps the DSP method achieve the superior performance.

Acknowledgements

References

[1] Q. Lai, T. Zhou, S. Khan, H. Sun, J. Shen, L. Shao, Weakly supervised visual saliency prediction, IEEE Transactions on Image Processing 31 (2022) 3111–3124.

[2] M. Jian, J. Wang, H. Yu, G.-G. Wang, Integrating object proposal with attention networks for video saliency detection, Information Sciences 576 (2021) 819–830.

[3] X. Fang, J. Zhu, X. Shao, H. Wang, Lc3net: Ladder context correlation complementary network for salient object detection, Knowledge-Based Systems 242 (2022) 108372.

[4] W. Ji, G. Yan, J. Li, Y. Piao, S. Yao, M. Zhang, L. Cheng, H. Lu, Dmra: Depth-induced multi-scale recurrent attention network for rgb-d saliency detection, IEEE Transactions on Image Processing 31 (2022) 2321–2336.

[5] P. Zhang, T. Zhuo, W. Huang, K. Chen, M. Kankanhalli, Online object tracking based on CNN with spatial-temporal saliency guided sampling, Neurocomputing 257 (2017) 115–127.

[6] M. Zhu, G. Hou, X. Chen, J. Xie, H. Lu, J. Che, Saliency-guided transformer network combined with local embedding for no-reference image quality assessment, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 1953–1962.

[7] S. Yang, Saliency-based image contrast enhancement with reversible data hiding, in: IEEE International Conference on Acoustics, Speech and Signal Processing, 2020, pp. 2847–2851.

[8] S. Zhu, Z. Xu, Spatiotemporal visual saliency guided perceptual high efficiency video coding with neural network, Neurocomputing 275 (2018) 511–522.

[9] L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (11) (1998) 1254–1259.

[10] N. Bruce, J. Tsotsos, Attention based on information maximization, Journal of Vision 7 (9) (2007) 950–950.

[11] D. Gao, N. Vasconcelos, Decision-theoretic saliency: computational principles, biological plausibility, and implications for neurophysiology and psychophysics, Neural Computation 21 (1) (2009) 239–271.

[12] X. Hou, L. Zhang, Saliency detection: A spectral residual approach, in: 2007 IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–8.

[13] W. Wang, J. Shen, Deep visual attention prediction, IEEE Transactions on Image Processing 27 (5) (2017) 2368–2378.

[14] R. Cong, J. Lei, H. Fu, M.-M. Cheng, W. Lin, Q. Huang, Review of visual saliency detection with comprehensive information, IEEE Transactions on circuits and Systems for Video Technology 29 (10) (2018) 2941–2959.

[15] T. Ma, W. Tian, Y. Xie, Multi-level knowledge distillation for low-resolution object detection and facial expression recognition, Knowledge-Based Systems (2022) 108136.

[16] L. Cheng, P. Fang, Y. Liang, L. Zhang, C. Shen, H. Wang, Tsgb: Target-selective gradient backprop for probing cnn visual saliency, IEEE Transactions on Image Processing 31 (2022) 2529–2540.

[17] B. Wang, X. Hu, C. Zhang, P. Li, S. Y. Philip, Hierarchical gan-tree and bi-directional capsules for multi-label image classification, Knowledge-Based Systems 238 (2022) 107882.

[18] K. R. Mopuri, U. Garg, R. V. Babu, Cnn fixations: an unraveling approach to visualize the discriminative image regions, IEEE Transactions on Image Processing 28 (5) (2018) 2116–2125.

[19] Y. Xu, S. Gao, J. Wu, N. Li, J. Yu, Personalized saliency and its prediction, IEEE Transactions on Pattern Analysis and Machine Intelligence 41 (12) (2018) 2975–2989.

[20] 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.

[21] K. Zhang, Z. Chen, S. Liu, A spatial-temporal recurrent neural network for video saliency prediction, IEEE Transactions on Image Processing 30 (2020) 572–587.

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

[23] D. Hu, X. Li, L. Mou, P. Jin, D. Chen, L. Jing, X. Zhu, D. Dou, Cross-task transfer for geotagged audiovisual aerial scene recognition, in: Proceedings of the European Conference on Computer Vision, 2020, pp. 68–84.

[24] Q. Song, B. Sun, S. Li, Multimodal sparse transformer network for audio-visual speech recognition, IEEE Transactions on Neural Networks and Learning Systems (2022).

[25] R. Qian, H. D. Di Hu, M. Wu, N. Xu, W. Lin, Multiple sound sources localization from coarse to fine, in: Proceedings of the European Conference on Computer Vision, 2020, pp. 292–308.

[26] Y. Aytar, C. Vondrick, A. Torralba, Soundnet: learning sound representations from unlabeled video, in: Advances in International Conference on Neural Information Processing Systems, 2016, pp. 892–900.

[27] K. Hara, H. Kataoka, Y. Satoh, Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2018, pp. 6546–6555.

[28] N. Liu, N. Zhang, K. Wan, L. Shao, J. Han, Visual saliency transformer, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 4722–4732.

[29] W. Wang, J. Shen, H. Ling, A deep network solution for attention and aesthetics aware photo cropping, IEEE Transactions on Pattern Analysis and Machine Intelligence 41 (7) (2018) 1531–1544.

[30] Y. Zhang, M. Jiang, Q. Zhao, Saliency prediction with external knowledge, in: Proceedings of the IEEE Winter Conference on Applications of Computer Vision, 2021, pp. 484–493.

[31] A. Kroner, M. Senden, K. Driessens, R. Goebel, Contextual encoder–decoder network for visual saliency prediction, Neural Networks 129 (2020) 261–270.

[32] J. Zhang, D.-P. Fan, Y. Dai, S. Anwar, F. Saleh, S. Aliakbarian, N. Barnes, Uncertainty inspired rgb-d saliency detection, IEEE Transactions on Pattern Analysis and Machine Intelligence (2021).

[33] T. Judd, K. Ehinger, F. Durand, A. Torralba, Learning to predict where humans look, in: 2009 IEEE 12th International Conference on Computer Vision, 2009, pp. 2106–2113.

[34] F. Perazzi, P. Krähenbühl, Y. Pritch, A. Hornung, Saliency filters: Contrast based filtering for salient region detection, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 733–740.

[35] Q. Wang, Y. Yuan, P. Yan, Visual saliency by selective contrast, IEEE Transactions on Circuits and Systems for Video Technology 23 (7) (2012) 1150–1155.

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

[37] M. Jiang, S. Huang, J. Duan, Q. Zhao, SALICON: Saliency in context, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1072–1080.

[38] S. He, H. R. Tavakoli, A. Borji, Y. Mi, N. Pugeault, Understanding and visualizing deep visual saliency models, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 10206–10215.

[39] W. Wang, J. Shen, Deep visual attention prediction, IEEE Transactions on Image Processing 27 (5) (2018) 2368–2378.

[40] K. Sss, K. Ayush, R. V. Babu, DeepFix: A fully convolutional neural network for predicting human eye fixations., IEEE Transactions on Image Processing 26 (9) (2017) 4446–4456.

[41] N. Liu, J. Han, T. Liu, X. Li, Learning to predict eye fixations via multiresolution convolutional neural networks, IEEE Transactions on Neural Networks and Learning Systems 29 (2) (2018) 392–404.

[42] Y. Yuan, H. Ning, X. Lu, Bio-inspired representation learning for visual attention prediction, IEEE Transactions on Cybernetics 51 (7) (2021) 3562–3575.

[43] M. Kummerer, T. S. A. Wallis, L. A. Gatys, M. Bethge, Understanding low- and high-level contributions to fixation prediction, in: IEEE International Conference on Computer Vision, 2017, pp. 4799–4808.

[44] X. Li, M. Chen, F. Nie, Q. Wang, A multiview-based parameter free framework for group detection, in: Thirty-First AAAI Conference on Artificial Intelligence, 2017.

[45] X. Li, B. Zhao, Video distillation, SCIENCE CHINA Information Sciences 51 (5) (2021) 695–734.

[46] L. Jiang, M. Xu, T. Liu, M. Qiao, Z. Wang, Deepvs: A deep learning based video saliency prediction approach, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 602–617.

[47] C. Bak, A. Kocak, E. Erdem, A. Erdem, Spatio-temporal saliency networks for dynamic saliency prediction, IEEE Transactions on Multimedia 20 (7) (2017) 1688–1698.

[48] S. Gorji, J. J. Clark, Going from image to video saliency: Augmenting image salience with dynamic attentional push, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7501–7511.

[49] M. Sun, Z. Zhou, Q. Hu, Z. Wang, J. Jiang, Sg-fcn: A motion and memory-based deep learning model for video saliency detection, IEEE Transactions on Cybernetics 49 (8) (2018) 2900–2911.

[50] Q. Lai, W. Wang, H. Sun, J. Shen, Video saliency prediction using spatiotemporal residual attentive networks, IEEE Transactions on Image Processing 29 (2019) 1113–1126.

[51] 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, Vol. 34, 2020, pp. 12410–12417.

[52] Z. Wang, Z. Zhou, H. Lu, Q. Hu, J. Jiang, Video saliency prediction via joint discrimination and local consistency, IEEE Transactions on Cybernetics (2020).

[53] J. Chen, H. Song, K. Zhang, B. Liu, Q. Liu, Video saliency prediction using enhanced spatiotemporal alignment network, Pattern Recognition 109 (2021) 107615.

[54] X. Min, G. Zhai, K. Gu, X. Yang, Fixation prediction through multimodal analysis, ACM Transactions on Multimedia Computing, Communications, and Applications 13 (1) (2016) 1–23.

[55] N. Sidaty, M.-C. Larabi, A. Saadane, Toward an audiovisual attention model for multimodal video content, Neurocomputing 259 (2017) 94–111.

[56] 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 Transactions on Image Processing 29 (2020) 3805–3819.

[57] D. Zhu, D. Zhao, X. Min, T. Han, Q. Zhou, S. Yu, Y. Chen, G. Zhai, X. Yang, Lavs: A lightweight audio-visual saliency prediction model, in: 2021 IEEE International Conference on Multimedia and Expo (ICME), 2021, pp. 1–6.

[58] D. Hu, R. Qian, M. Jiang, X. Tan, S. Wen, E. Ding, W. Lin, D. Dou, Discriminative sounding objects localization via self-supervised audiovisual matching, Advances in Neural Information Processing Systems 33 (2020).

[59] T. Afouras, A. Owens, J. S. Chung, A. Zisserman, Self-supervised learning of audio-visual objects from video, in: European Conference on Computer Vision, 2020, pp. 208–224.

[60] A. Tsiami, P. Koutras, P. Maragos, Stavis: Spatio-temporal audiovisual saliency network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 4766–4776.

[61] J. Chen, Q. Li, H. Ling, D. Ren, P. Duan, Audiovisual saliency prediction via deep learning, Neurocomputing 428 (2021) 248–258.

[62] H. R. Tavakoli, A. Borji, E. Rahtu, J. Kannala, Dave: A deep audio-visual embedding for dynamic saliency prediction, arXiv preprint arXiv:1905.10693 (2019).

[63] A. Senocak, T.-H. Oh, J. Kim, M.-H. Yang, I. S. Kweon, Learning to localize sound source in visual scenes, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4358–4366.

[64] Y. Chen, X. Lu, S. Wang, Deep cross-modal image–voice retrieval in remote sensing, IEEE Transactions on Geoscience and Remote Sensing 58 (10) (2020) 7049–7061.

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

[66] Y. Zhang, Y. Yuan, Y. Feng, X. Lu, Hierarchical and robust convolutional neural network for very high-resolution remote sensing object detection, IEEE Transactions on Geoscience and Remote Sensing 57 (8) (2019) 5535–5548.

[67] N. Komodakis, S. Zagoruyko, Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer, in: International Conference on Learning Representations, 2017.

[68] S. Woo, J. Park, J.-Y. Lee, I. S. Kweon, Cbam: Convolutional block attention module, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 3–19.

[69] M. Cornia, L. Baraldi, G. Serra, R. Cucchiara, Predicting human eye fixations via an lstm-based saliency attentive model, IEEE Transactions on Image Processing 27 (10) (2018) 5142–5154.

[70] R. Droste, J. Jiao, J. A. Noble, Unified image and video saliency modeling, in: European Conference on Computer Vision, 2020, pp. 419–435.

[71] A. Coutrot, N. Guyader, Multimodal saliency models for videos, in: From Human Attention to Computational Attention, 2016, pp. 291–304.

[72] A. Coutrot, N. Guyader, How saliency, faces, and sound influence gaze in dynamic social scenes, Journal of Vision 14 (8) (2014) 1–17.

[73] P. K. Mital, T. J. Smith, R. L. Hill, J. M. Henderson, Clustering of gaze during dynamic scene viewing is predicted by motion, Cognitive Computation 3 (1) (2011) 5–24.

[74] P. Koutras, P. Maragos, A perceptually based spatio-temporal computational framework for visual saliency estimation, Signal Processing: Image Communication 38 (2015) 15–31.

[75] A. Tsiami, P. Koutras, A. Katsamanis, A. Vatakis, P. Maragos, A behaviorally inspired fusion approach for computational audiovisual saliency modeling, Signal Processing: Image Communication 76 (2019) 186–200.

[76] M. Gygli, H. Grabner, H. Riemenschneider, L. Van Gool, Creating summaries from user videos, in: European Conference on Computer Vision, 2014, pp. 505–520.

[77] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, F. Durand, What do different evaluation metrics tell us about saliency models?, IEEE Transactions on Pattern Analysis and Machine Intelligence 41 (3) (2018) 740–757.

[78] 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.

[79] 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, arXiv preprint arXiv:1701.01081 (2017).

[80] W. Wang, J. Shen, J. Xie, M.-M. Cheng, H. Ling, A. Borji, Revisiting video saliency prediction in the deep learning era, IEEE Transactions on Pattern Analysis and Machine Intelligence 43 (1) (2019) 220–237.

[81] K. Min, J. J. Corso, Tased-net: Temporally-aggregating spatial encoder-decoder network for video saliency detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 2394–2403.