Bio-Inspired Audio-Visual Cues Integration for Visual Attention Prediction

Yuan Yuan, Senior Member, IEEE, Hailong Ning, and Bin Zhao

Abstract—Visual Attention Prediction (VAP) methods simulate the human selective attention mechanism to perceive the scene, which is significant and imperative in many vision tasks. Most 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 simultaneously sensing these cues. Motivated by this, a bio-inspired audio-visual cues integration method is proposed for the VAP task, which explores the audio modality to better predict the visual attention map by assisting vision modality. The proposed method consists of three parts: 1) audio-visual encoding, 2) audio-visual location, and 3) multi-cues aggregation 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 multi-cues aggregation part is devised to adaptively aggregate audio-visual information and center-bias prior to generate the final visual attention map. Extensive experiments are conducted on six challenging audiovisual eye-tracking datasets, including DIEM, AVAD, Coutrot1, Coutrot2, SumMe, and ETMD, which shows significant superiority over state-of-the-art visual attention models.

Index Terms—Visual Attention Prediction, Bio-Inspired, Audio-Visual, Multi-Modal, Motion

I. INTRODUCTION

Visual attention prediction (VAP) task aims to automatically predict the most prominent area in the scene by simulating the human selective attention mechanism, which provides an effective idea 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. 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 visual attention map in this paper. Specifically, humans

A. 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 audio can significantly promote the understanding of the scene [19]–[22]. For example, Hu et al. [20] propose 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 VAP have not fully realized the potential contribution of audio information to the performance. They predict the visual attention map by only mining the information in the visual modality itself, while ignoring the latent effect of the accompanied audio information.

In the field of computer vision, the VAP 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 VAP methods and dynamic VAP methods. The static VAP 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 VAP 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 visual attention map. With the popularity of deep learning, a large number of researchers conduct the VAP task by mining high-level semantic information in images [13], [15]–[17]. For example, Wang et al. [13] obtain the hierarchically saliency information by extracting multi-scale features of images. The dynamic VAP 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. [18] design two parts to extract static spatial features and dynamic temporal features for VAP. In view of practical applications, this paper aims to investigate the VAP for the dynamic video.

This work was supported in part by the National Key R&D Program of China under Grant 2020YFB103902, in part by the National Science Fund for Distinguished Young Scholars under Grant 61825603, in part by the Key Program of National Natural Science Foundation of China under Grant 61632018, and in part by China Postdoctoral Science Foundation under Grant 2020TQ0236. (Corresponding author: Bin Zhao)

Manuscript received XX XX, XXXX.
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 VAP 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-priors, how to better integrate them. These inconsistencies will bring some interference and are the main challenges for the VAP.

Based on the above opportunities and challenges, a bio-inspired audio-visual cues integration method is proposed for the VAP task. The proposed method can be purposely leveraged to predict the visual attention map in videos by simultaneously integrating the audio-visual cues. Concretely, the proposed method is composed of three main parts: 1) audio-visual encoding, 2) audio-visual location, and 3) multi-cues aggregation 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 [23], and the visual branch employs a modified 3D ResNet-50 architecture [24]. 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 multi-cues aggregation part is introduced to adaptively aggregate multi-cues information, so as to generate the final visual attention 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.

B. Contributions

Generally, the main contributions of this paper are threefold:

• A bio-inspired audio-visual cues integration method is proposed, which can comprehensively consider the influence of audio-visual cues on visual attention prediction, and realize the role of auxiliary enhancement.

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

• A multi-cues attention aggregation module is designed, which can adaptively integrate the influences of multiple cues on visual attention.

C. Organization

The remaining sections of the paper are arranged as follows: In Section II, the related works are reviewed. In Section III, we describe the proposed method in details. In Section IV, the experimental results are shown and discussed. And in Section V, the conclusion is drawn.

II. RELATED WORKS

In recent years, numerous VAP methods have been proposed. According to the data type, the existing methods can be divided into two aspects, including static VAP methods and dynamic VAP methods. This section reviews the existing VAP methods in the following successively.

A. Static VAP Methods

Earlier works devote to investigate the VAP task based on static images [25–29]. These works are mostly on the basis of bottom-up visual attention mechanism [25], [26], [30], [31]. Itti et al. [9] firstly conduct the VAP 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. [31] consider 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 VAP have been proposed. Perazzi et al. [32] design a conceptually clear and intuitive VAP method based on the contrast feature. Wang et al. [33] leverage selective contrast, including color, texture, and location, to predict the salient regions of images. In addition, many researchers have realized the prediction of the visual attention map in static images 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 can not be well processed.

As a result, most researchers begin to conduct the VAP task by mining high-level semantic information in images with deep neural networks [13], [15]–[17]. Vig et al. [34] firstly propose utilizing deep neural networks for the VAP task. However, due to the insufficient training data, the performance is limited. To address this problem, Jiang et al. [35] build the SALICON dataset with plenty of natural images and the corresponding eye-tracking data. Based on the dataset, lots of works on VAP are developed by the follow-up researchers [36]. He et al. [36] explore the intrinsic 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 [27]–[29], [37]–[39]. Zhang et al. [28] propose to incorporate prior knowledge of semantic relationships so as to learn highlighted regions in images. Kroner et al. [29] develop an encoder-decoder structure to learn multiscale high-level visual features for VAP. Considering that both low-level contrast features and high-level semantic features are important for VAP, Yuan et al. [40] introduce a bio-inspired representation learning method to generate the visual attention map. Wang et al. [37] conduct the VAP by fusing features from multiple layers of VGG-16. Kruthivent et al. [41] consider the center-bias prior information and develop a computational method for VAP, which improved the predicted results considerably.
B. Dynamic VAP Methods

With the tremendous progress of data storage technology and mobile Internet technology, massive video data are soaring recently [42], [43]. In order to deal with the vast amounts of information, researchers pay more attention on dynamic VAP to predict the most valuable information in video data [44]–[51]. Jiang et al. [44] develop a object-to-motion convolutional neural network for predicting the intra-frame visual attention. Bak et al. [45] present a spatio-temporal saliency network for dynamic VAP. Gorji et al. [46] propose a multi-stream ConvLSTM structure with the attentional push effect of the scene actors and the photographer. To improve the performance of VAP models, Sun et al. [47] put forward a step-gained fully convolutional network by simultaneously considering motion and temporal information. By conducting multi-scale feature learning and spatiotemporal feature integration, Lai et al. [48] design a residual attentive learning network for dynamic VAP. Wu et al. [49] present a end-to-end neural network named SAC for dynamic VAP. 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 VAP, Wang et al. [50] introduce a dynamic saliency network on the basis of both global discriminations and local consistency. These methods have greatly promoted the progress of dynamic VAP. Nevertheless, the effect of the audio information accompanying the video is ignored when conducting VAP 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 visual attention map. Starting from application-specific, some researcher adopt the traditional signal processing techniques for locating the salient region in the scene [52]–[55]. For example, Min et al. [55] utilize 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. Qian et al. [22] propose 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. [56] introduce a two-stage learning framework with a self-supervised class-aware manner. Afouras et al. [57] develop a model using attention to transform a video into several discrete audio-visual objects. Tavakoli et al. [58] design a conceptually simple and effective audio-visual analysis method for dynamic saliency prediction. Tsiami et al. [59] propose a spatio-temporal audio-visual saliency network by combining both visual and auditory information.

This paper is dedicated to conduct the dynamic VAP by integrating the audio-visual cues, while alleviating existing inconsistencies between audio-visual modalities. The proposed method is bio-inspired since the sounding, moving, center, high-level semantic objects are more attractive, and these cues are considered simultaneously by the proposed method.

III. THE PROPOSED METHOD

In this work, a bio-inspired audio-visual cues integration method is proposed to generate visual attention map. As is shown in Fig. [1] the proposed method is composed of
three main parts: 1) audio-visual encoding, 2) audio-visual location, and 3) multi-cues aggregation 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 multi-cues aggregation 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 visual attention map. These parts are expatiated successively as follows.

A. 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.

1) Audio Encoding: In the VAP task based on audio-visual analysis, it is important to obtain the semantic concept of audio rather than low-level signal [60]. To this end, the audio signals are represented by using convolutional neural networks (CNNs). Specifically, we follow the previous work [59] 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 [23] and a temporal max-pooling layer is applied for encoding high-level semantic features. Formulically, the process of audio encoding can be written as:

\[
S_A = \mathcal{F}_A(\mathbf{X}_A; \theta_A),
\]

where \(\mathbf{X}_A\) and \(S_A\) 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. \(\theta_A\) is the parameter during the process of audio encoding.

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 [24], which is proposed initially for action recognition, is employed as the backbone to encode the spatio-temporal features. Lots of works [61–63] 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 VAP in this work. As is shown in [1], the spatio-temporal visual encoding branch adopts the first 4 ResNet convolutional blocks to provide the outputs \(S_{V}^1, S_{V}^2, S_{V}^3, S_{V}^4\), which contain different spatial and temporal information. The process of spatio-temporal visual encoding can be written as:

\[
S_{V}^m = \mathcal{F}_V(\mathbf{X}_V; \theta_V^m),
\]

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

B. Audio-Visual Location

In this part, an audio-visual location part is devised to locate the sounding object by exploiting the consistency in a sharing latent space of the audio and visual modality. 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 \(S_{V}^4\), 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 \(S_{V}^4\). For simplicity, the global representation is denoted as a reshaped matrix form \(\mathbf{V} = [\mathbf{v}_1; \cdots; \mathbf{v}_B] \in \mathbb{R}^{B \times D_v}\). 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 \(S_A\), 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 Fig. [1]) to locate the sounding object and generate a audio-aware visual attention map. It is achieved as follows:

\[
a_b = \langle \mathbf{W}_1 \mathbf{v}_b, \mathbf{W}_2 \mathbf{h}_A \rangle,
\]

\[
a_b = \frac{\exp(a_b)}{\sum_{b=1}^{B} \exp(a_b)},
\]

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 matrices. \(a_b\) stands for the \(b\)-th element in the sounding map \(\alpha\) which can be interpreted as the probability of location related to the audio context. Further, the audio-aware visual attention map \(\mathbf{F}_{audio}\) 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. After obtain the sounding map, we fourthly compute the representative context vector \(\mathbf{h}_c\) to interact the sounding map with the global representation in visual branch at the sound source location. The process is achieved by:

\[
\mathbf{h}_c = \sum_{b=1}^{B} a_b \mathbf{v}_b.
\]

\(^1\)https://distill.pub/2016/deconv-checkerboard/
Then, the representative context vector $h_z$ is transformed to a spatio-temporal visual feature $\hat{v}$ with two fully connected (FC) layers with ReLU activation. Finally, a location loss is imposed on $\hat{v}$ and $S_A$ to learn the features to share latent space of the audio and visual modality. Here, the Euclidean distance loss function is employed as the location loss.

### C. Multi-Cues 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, a bio-inspired audio-visual cues integration method is proposed for the VAP task. We have obtained the audio-aware visual attention map according to the audio-visual location part. Thus, this subsection is to describe how to model the moving, center, and high-level semantic objects for the VAP task. Successively, a multi-cues aggregation module is introduced to integrate the influence of different cues and generate the final visual attention map.

1) **Moving and High-Level Semantic Objects Prediction:** In order to model the influence of moving and semantic objects for VAP, 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 Fig. 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 $S_{V,T}^m$. Secondly, the first and last frames are removed, respectively, leading to $S_{V,T,-1}^m$ and $S_{V,T,0}^m$. Thirdly, the temporal attention is computed by conducting the frame-wise similarity between $S_{V,T,-1}^m$ and $S_{V,T,0}^m$ so as to capture the temporally moving information. Specifically, the frame-wise similarity can be implemented as:

$$M_T^m = \sum_{t=1}^{t=T-1} (1 - (S_{V,T,0}^m - S_{V,T,-1}^m)),$$  \hspace{1cm} (6)

where $M_T^m$ represent the motion feature from the $m$-th ResNet convolutional block. $t$ indexes the frames. Afterwards, the motion feature is further processed with a $1 \times 1$ convolution and a resize-convolution operation to generate a motion-aware attention map $F_{\text{motion}}$.

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 Fig. 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 $S_{V,C}^m$. 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 [64], [65]. Thirdly, a convolution layer with filter size of $7 \times 7$ is conducted for generating a spatial weight matrix $M_{\text{weight}}^m$. Fourthly, by employing an element-wise multiplication between $S_{V,C}^m$ and the spatial weight matrix $M_{\text{weight}}^m$, the feature $S_{V,T}^m$ with high-level semantic information is learned. Finally, $S_{V,T}^m$ is further processed with the $1 \times 1$ convolution and resize-convolution operation to generate a high-level semantic-aware attention map $F_{\text{semantic}}^m$.

2) **Center-Bias Prior:** According to previous studies [28], [66], [67], 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 [40]. Specifically, the center-bias prior-aware map $F_{\text{center}}$ is generated using a Gaussian function as follows:

$$F_{\text{center}} = \frac{1}{2\pi\sigma_x\sigma_y} \exp \left(-\frac{(x-x_0)^2}{2\sigma_x^2} + \frac{(y-y_0)^2}{2\sigma_y^2} \right),$$  \hspace{1cm} (7)

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

3) **Multi-Cues Aggregation:** We have computed the audio-aware attention map, the motion-aware attention map, the high-level semantic-aware attention map, and the center-bias prior-aware map. They can express the visual attention driven by different cues. As a result, it is quite essential to integrate them for generating the final visual attention 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 Fig. 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 module is composed of three main steps. Firstly, the concatenated feature $M_{\text{conc}} = [F_{\text{audio}}; F_{\text{motion}}; F_{\text{semantic}}; F_{\text{center}}]$ is processed as the global response context $g$ by:
The proposed multi-cues aggregation module.

\[ g = \text{global}(M_{\text{conc}}; W_3) \]
\[ = \sigma(B(PWC(\sigma(B(PWC(GAP(M_{\text{conc}}))))))) \],

(8)

where global denotes the global response mapping function. \( W_3 \) is the to-be-learned parameter. \( \sigma \) represents the Sigmoid function. \( B \) means the Batch Normalization (BP) operation. PWC stands for the Point-Wise Convolution (PWC), which is chosen for its lightweight. GAP represents the global average pooling.

In parallel, the local response context \( L \) can be computed by:

\[ L = \text{local}(M_{\text{conc}}; W_4) \]
\[ = B(PWC(\delta(B(PWC(M_{\text{conc}}))))) \],

(9)

where local denotes the local response mapping function. \( W_4 \) is the to-be-learned parameter. \( \delta \) represents the ReLU function.

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

\[ F_{\text{fusion}} = g \odot L, \]

(10)

where \( F_{\text{fusion}} \) represents the fused feature by considering different cues, and \( \odot \) means the channel-wise multiplication.

Finally, based on the fused feature \( F_{\text{fusion}} \), the final visual attention map \( F_{\text{map}} \) is computed by a Readout Network, which is composed of 3 successive \( 1 \times 1 \) convolution layers.

**D. Optimizing Strategy**

In order to obtain the final visual attention map, we aggregate the visual attention map driven by cues of sounding, moving, center, high-level semantic objects. Let \( Y_{\text{sal}} \) represent the ground-truth fixation map obtained by the eye-tracking data. During audio-visual location process, the objective function is defined as:

\[ L_A = KL(F_{\text{audio}}, Y_{\text{sal}}) + \| (\hat{v}, S^A) \|_2 \]
\[ = Y_{\text{sal}} \log \left( \frac{Y_{\text{sal}}}{F_{\text{audio}} + \epsilon} \right) + \| (\hat{v}, S^A) \|_2, \]

(11)

where \( KL(\cdot, \cdot) \) stands for the Kullback–Leibler divergence between the two distributions. \( \| \cdot \|_2 \) is the Euclidean distance. \( \epsilon \) indicates the regularization constant to avoid the NaN value in the loss.

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

\[ L_{MS} = KL(F_{\text{motion}}, Y_{\text{sal}}) + KL(F_{\text{semantic}}, Y_{\text{sal}}) \]
\[ = Y_{\text{sal}} \log \left( \frac{Y_{\text{sal}}}{F_{\text{motion}} + \epsilon} \right) + Y_{\text{sal}} \log \left( \frac{Y_{\text{sal}}}{F_{\text{semantic}} + \epsilon} \right). \]

(12)

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

\[ L_{\text{fuse}} = KL(F_{\text{map}}, Y_{\text{sal}}) \]
\[ = Y_{\text{sal}} \log \left( \frac{Y_{\text{sal}}}{F_{\text{map}} + \epsilon} \right). \]

(13)

Ultimately, the final loss \( L_{\text{final}} \) of training all parameters can be jointly combined by the losses \( L_A, L_{MS}, \) and \( L_{\text{fuse}} \), as follows.

\[ L_{\text{final}} = w_1 L_A + w_2 L_{MS} + w_3 L_{\text{fuse}}, \]

(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 \( X^v \), and the corresponding audio data \( X^a \);

Testing video frames \( X'^v \), and the corresponding audio data \( X'^a \).

**Output:**

Testing visual attention map \( F_{\text{map}}'^a \);

All the to-be-learned parameters \( W \).

**Initialization:**

The parameter \( \theta_A \) in the audio encoding branch is initialized from the origin SoundNet [23]. The parameter \( \theta_V \) in the spatio-temporal visual encoding branch is initialized from the origin 3D ResNet-50 [24]. The remaining weights are randomly initialized by truncated normal distribution.

**Repeat:**

1: Calculate the high-level semantic feature \( S^A \) and spatio-temporal feature \( S^V \) according to Eq. 1 and Eq. 2, respectively;

2: Generate the audio-aware visual attention map \( F_{\text{audio}} \) according to Eq. 3 and Eq. 4;

3: Calculate the motion-aware attention map \( F_{\text{motion}} \), the high-level semantic-aware attention map \( F_{\text{semantic}} \), and center-bias prior-aware map \( F_{\text{center}} \) based on Section III-C1 and Section III-C2;

4: Generate the final attention map \( F_{\text{map}} \) based on Section III-C3;

5: Compute the final loss \( L_{\text{final}} \) according to Eq. 14;

6: Update all the parameters by utilizing Adam optimizer.

**Until:** A fixed number of iterations.

7: Generate the testing visual attention map \( F_{\text{map}}'^a \).

**Return:** \( F_{\text{map}}'^a, W \)

Based on the final loss \( L_{\text{final}} \), the proposed method can be optimized as follows. The parameter \( \theta_A \) in the audio
encoding part are initialized from the origin SoundNet [23], which is pretrained on. The parameter $\theta_V$ in the spatio-temporal visual encoding part is initialized from the origin 3D ResNet-50 [24], 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 $X_V$ and the corresponding audio data $X_A$ are processed as high-level semantic feature $S_A$ and spatio-temporal feature $S_V$ with the audio-visual encoding part. Secondly, the high-level semantic feature $S_A$ and spatio-temporal feature $S_V$ are combined to locate the sounding object to generate the audio-aware visual attention map $F_{\text{audio}}$. Thirdly, the motion-aware attention map $F_{\text{motion}}$, the high-level semantic-aware attention map $F_{\text{semantic}}$, and center-bias prior-aware map $F_{\text{center}}$ are computed based on Section III-C1 and Section III-C2. Fourthly, the final attention map $F_{\text{map}}$ are inferred based on $F_{\text{audio}}$, $F_{\text{motion}}$, $F_{\text{semantic}}$, and $F_{\text{center}}$. Finally, we compute the final loss $L_{\text{final}}$ based on the generated final attention map $F_{\text{map}}$ and the ground-truth fixation map $Y_{\text{sal}}$, and update all the parameters $W$ by minimizing $L_{\text{final}}$ with Adam optimizer. Once the training epoch reaches 50, the training process is terminated. Afterwords, the parameter $W$ is utilized to infer the testing visual attention map $F_{\text{te}_{\text{map}}}$. 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.

IV. 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.

A. Setup

1) Datasets: The proposed method is trained and evaluated on AVAD [53], Coutrot1 [68], Coutrot2 [68], DIEM [70], ETMD [71], ETMD dataset includes 12 video clips from several Hollywood movies. The eye-tracking data are annotated by 10 different people. The SumMe dataset [72], [73] 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 [59], we adopt the same data partitioning for training 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.

2) Evaluation Metrics: In order to measure the consistency between the predicted visual attention map and the groundtruth fixation map, 5 widely-used evaluation metrics for V AP are employed [74]. 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 visual attention 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 visual attention map. The SIM measures the similarity between the predicted visual attention map and groundtruth fixation map. The 5 evaluation metrics provide a comprehensive assessment for VAP.

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 [59], the data augmentation is also employed for random generation of training samples. The implementation adopts the 3D ResNet-50 [24] as backbone for encoding spatio-temporal visual features, and applies SoundNet [23] as backbone for encoding high-level audio semantic features. The Gaussian kernel for generating center-bias prior map is with size of $7 \times 7$. The tradeoff coefficients $w_1$, $w_2$, and $w_3$ in Eq. 14 are all selected as 1. 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 visual attention map of each frame. The experiment is conducted by the Pytorch library and on the PC with a TITAN RTX GPU and 24G RAM.

B. 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 VAP, 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 in a sharing latent space between the audio and visual modality.
- **Concatenate Fusion**: In order to integrate the multi-cues maps and generate the final visual attention map, we
TABLE I
THE ABLATION STUDY ON SIX BENCHMARK DATASETS.

| Methods               | AVAD          |            |            | DIEM          |            |            | SumMe        |
|-----------------------|---------------|------------|------------|---------------|------------|------------|--------------|
|                       | CC            | NSS        | sAUC       | SIM           | CC         | NSS        | sAUC         | SIM           |
| Visual Model          | 0.5714        | 3.11       | 0.8928     | 0.5740        | 0.4635     |            | 0.4558       | 2.03          | 0.8527        | 0.5562       | 0.3630       | 0.4260       | 3.26       | 0.9208       | 0.6444       | 0.3971       |
| AV Inner-Product      | 0.5751        | 3.14       | 0.9004     | 0.5827        | 0.4572     |            | 0.4608       | 2.16          | 0.8604        | 0.5681       | 0.3744       | 0.6748       | 4.68       | 0.9370       | 0.6727       | 0.4283       |
| Concatenate Fusion    | 0.5826        | 3.32       | 0.9081     | 0.5944        | 0.4680     |            | 0.4714       | 2.29          | 0.8702        | 0.5774       | 0.3853       | 0.7034       | 4.91       | 0.9435       | 0.6911       | 0.4824       |
| Proposed (w/o SA)     | 0.5907        | 3.36       | 0.9127     | 0.6014        | 0.4752     |            | 0.4807       | 2.35          | 0.8696        | 0.5842       | 0.3967       | 0.7221       | 5.09       | 0.9466       | 0.7028       | 0.4960       |
| Proposed (w/o TA)     | 0.6125        | 3.44       | 0.9204     | 0.6145        | 0.4783     |            | 0.4852       | 2.36          | 0.8704        | 0.5861       | 0.4017       | 0.7283       | 5.14       | 0.9473       | 0.7135       | 0.5048       |
| Proposed              | 0.6262        | 3.57       | 0.9251     | 0.6203        | 0.4820     |            | 0.4985       | 2.44          | 0.8798        | 0.6042       | 0.4154       | 0.7481       | 5.45       | 0.9537       | 0.7294       | 0.5266       |

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

directly concatenate them and further readout, rather than by utilize 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 I 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 dynamic VAP. 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 from 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. And the sAUC metric drops nearly 8% on the SumMe dataset. The results demonstrate the important role of audio information for the dynamic VAP 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, Fig. 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 visual attention 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 model locates the non-sounding person, which is no salient. This further demonstrates 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 VAP methods. For the observation from TABLE I, we can find that when the devised audio-visual location is replaced with inner-product operation, the performance for VAP 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 I, we can find the great differences between the results from Concatenate Fusion method and the proposed method. It mainly because the Concatenate Fusion method can not effectively integrate the influences of multiple cues on visual attention. 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 VAP. 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. It is mainly because the spatial attention can capture the temporally moving information by exploiting the frame-wise similarity. The results verify the importance of the spatial attention for capturing the higher-level semantic information.

5) The temporal attention lift the performance of dynamic VAP methods by modeling motion information. The comparison results between Proposed (w/o TA) and the proposed method on TABLE I 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 dynamic VAP methods.

C. 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 VAP 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 visual attention map, while do not consider the temporal information among frames. The comparison spatial methods include DeepNet [75], DVA [13], SAM [66], and SalGAN [76] methods. DeepNet [75] addresses the VAP methods. These comparison methods are comprehensive, including 4 spatial methods and 5 spatio-temporal methods.

2) The spatio-temporal methods process each frame independently to generate the visual attention map, while do not consider the temporal information among frames. The comparison spatio-temporal methods include ACLNet [77], DeepVS [44], TASED [78], and STA ViS [59] methods. ACLNet [77] employs the CNN-LSTM architecture for dynamic VAP with a supervised attention mechanism. DeepVS [44] develops an object-to-motion convolutional neural network for estimating the intra-frame saliency. TASED [78] exploits 3D fully-convolutional network architecture to generate the visual attention map of each frame by considering several past frames. STA ViS [59] combines both visual and auditory information for VAP in videos.

On the on hand, we compare the proposed method with 8 state-of-the-art VAP methods qualitatively on six benchmark datasets with audio-visual eye-tracking data, including AVAD [53], Coutrot1 [68], Coutrot2 [68], [69], DIEM [70], ETMD [71], [72] and SumMe [72], [73] datasets. TABLE I 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 VAP methods, such as DeepNet [75], DVA [13], SAM [66], and SalGAN [76], by a substantial margin. The success attributes to the proposed method capturing the temporal information, which is quite important for dynamic VAP. Compared with the spatio-temporal VAP methods, e.g., ACLNet [77], DeepVS [44], TASED [78], and STA ViS [59], the proposed method also exhibits better performance. It is mainly because the proposed method can adaptively integrate multiple cues, which are essential for VAP. More specifically, even though the STA ViS method also adopts audio-visual information for VAP,
the proposed method surpasses it on the prediction of visual attention map. This because we integrate more information affecting VAP, design more effective audio-visual location part for locating the sounding objects, and propose multi-cues aggregation module to adaptively integrate the influences of multiple cues on visual attention. In addition, the higher CC and SIM values show that the generated visual attention 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 Fig. 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 adaptively integrates multiple cues for VAP. 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 VAP, and the superiority of audio-visual location part for locating sounding objects and the multi-cues aggregation module for adaptively integrating multi-cues information.

V. CONCLUSIONS

In this paper, we propose a bio-inspired audio-visual cues integration method for visual attention 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 multi-cues aggregation 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 in a sharing latent space of the audio and visual modality. The multi-cues aggregation part captures the moving, center, and high-level semantic information, and adaptively integrates them with the sounding information to generate the final visual attention map. The experimental results on six benchmark datasets have demonstrated that: 1) the audio information contributes to the VAP task significantly; 2) the consistency in a sharing latent space of the audio and visual modality enhances the performance of sounding object location; 3) the adaptively aggregation of multi-cues helps the VAP method achieve the superior performance.
