CASP-Net: Rethinking Video Saliency Prediction from an Audio-Visual Consistency Perceptual Perspective

Junwen Xiong[1*], Ganglai Wang[1*], Peng Zhang[1,2*†], Wei Huang[3], Yufei Zha[1,2], Guangtao Zhai[4]
[1]Northwestern Polytechnical University [2]Ningbo Institute of Northwestern Polytechnical University
[3]Nanchang University [4]Shanghai Jiao Tong University

2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) | 979-8-3503-0129-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/CVPR52729.2023.00623

Abstract

Incorporating the audio stream enables Video Saliency Prediction (VSP) to imitate the selective attention mechanism of human brain. By focusing on the benefits of joint auditory and visual information, most VSP methods are capable of exploiting semantic correlation between vision and audio modalities but ignoring the negative effects due to the temporal inconsistency of audio-visual intrinsics. Inspired by the biological inconsistency-correction within multi-sensory information, in this study, a consistency-aware audio-visual saliency prediction network (CASP-Net) is proposed, which takes a comprehensive consideration of the audio-visual semantic interaction and consistent perception. In addition a two-stream encoder for elegant association between video frames and corresponding sound source, a novel consistency-aware predictive coding is also designed to improve the consistency within audio and visual representations iteratively. To further aggregate the multi-scale audio-visual information, a saliency decoder is introduced for the final saliency map generation. Substantial experiments demonstrate that the proposed CASP-Net outperforms the other state-of-the-art methods on six challenging audio-visual eye-tracking datasets. For a demo of our system please see our project webpage.

1. Introduction

The task of saliency prediction is to automatically estimate the most prominent area in a scenario by simulating human selective attention. It has been extended to an alternative way to extract the most valuable information from a massive of data, which serves wide applications such as robotic camera control [7], video captioning [35], motion tracking [30], image quality evaluation [50] and video compression [51], *etc*.

In recent years, a lot of saliency prediction works have been developed by their increasing attention [4, 15, 41–43, 49]. According to different data types, these studies can be categorized into Image Saliency Prediction (ISP) and Video Saliency Prediction (VSP). The ISP investigates how to combine the low-level heuristic characteristics (*e.g*., colour, texture and luminance) with high-level semantic image attributes to predict prominent areas in the scene [15, 41, 42]. Differently, VSP exploits how to apply the spatio-temporal structure information in videos, and benefits the perception and identification of dynamic scenes [4, 49].

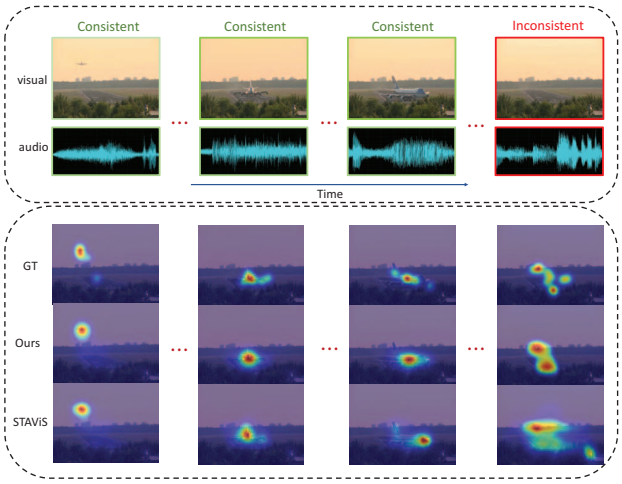

Figure 1. The example figure shows the saliency results of our model compared to STAViS [39] in audio and video temporal sequences. In the last time segment, the audio information that occurs in the event is inconsistent with the visual information. Our method can cope with such challenge by automatically learning to align the audio-visual features. The results of STAViS, however, show that it is incapable to address the problem of audio-visual inconsistency. *GT* denotes ground truth.

From the view of data modalities, the vision and audio present the video content from different sensing, which complement each other to enhance the perception. Based on multi-modal data, more recent studies have that audio information can significantly improve the understanding of the video semantics [33, 37, 39]. Min *et al*. [33] conduct a cross-modal kernel canonical correlation analysis (CCA)

*Indicates equal contributions.
†Corresponding author: zh0036ng@nwpu.edu.cn.

979-8-3503-0129-8/23/$31.00 ©2023 IEEE
DOI 10.1109/CVPR52729.2023.00623

by exploring audio-visual correspondence clues, and effectively enhance the video-level saliency prediction accuracy. Tsiami *et al*. [39] propose a deep model by combining spatio-temporal visual and auditory information to address the video saliency estimation efficiently. Nevertheless, these works heavily depend on temporal consistency of visual and audio information, and thus may suffer an unexpected degradation in practical scenarios, where such consistency cannot be satisfied as shown in Figure 1.

Temporal inconsistency commonly exists in real-life videos because realistic visual scenarios usually contain multiple sound sources, which may come from on-screen (*e.g*., dialogue in a talk show), or from off-screen (*e.g*., narration in a movie). Without understanding the complex scenario components, simply performing audio-visual consistency learning would result in an irrelevant semantic matching. A promising solution to this challenge is motivated by the study of neuroscience [18, 36], which explains how our brain minimizes the matching errors within multisensory data using both iterative inference and learning, and also inspired the **C**onsistency-aware **A**udio-visual **S**aliency **P**rediction network *CASP-Net* of this study.

By substantially exploring the latent semantic correlations of cross-modal signals, in *CASP-Net*, the potential temporal inconsistency between different modalities can be corrected as well. In addition, a two-stream network is also introduced to elegantly associate video frames with the corresponding sound source, which is able to achieve semantic similarities between audio and visual features by cross-modal interaction. To further reason the coherent visual and audio content in an iterative feedback manner, a consistency-aware predictive coding (CPC) module is designed. Subsequently, a saliency decoder (SalDecoder) is proposed to aggregate the multi-scale audio-visual information from all previous decoder's blocks and to generate the final saliency map. The main contributions in this work can be summarized as follows:

(1) A novel audio-visual saliency prediction model is proposed by comprehensively considering the functionalities of audio-visual semantic interaction and consistent perception. (2) A consistency-aware predictive coding module is designed to improve the consistency within audio and visual representations iteratively. (3) Solid experiments have been conducted on six audio-visual eye-tracking datasets, which demonstrate a superior performance of the proposed method in comparison to the other state-of-the-art works.

2. Related Work

2.1. Video Saliency Prediction

For video saliency prediction, different strategies of modeling temporal motion information have been proposed to estimate the saliency maps over consecutive frames [4, 26, 31]. Bak *et al*. [4] propose a two-stream spatio-temporal network to process video frames with optical flow maps, simultaneously. The Long-Short Term Memory (LSTM) [24] and Gated Recurrent Unit (GRU) [14] have also been incorporated into the video saliency prediction, and subsequently Wang *et al*. [43] propose to combine the ConvLSTM with dynamic attention mechanism into a network to further enhance the prediction performance. Similarly to model long-term temporal characteristics, Lai *et al*. [29] propose STRA-Net based on a lightweight convGRU. After Min *et al*. [31] adopt a S3D model [45] to build TASED-Net with 3D convolutions, the training paradigm of 3D convolution has been widely used in the VSP task. Bellitto *et al*. [5] design a 3D fully convolutional architecture to obtain multi-scale saliency instances for the combination of output saliency maps. Jain *et al*. [26] adopt a 3D encoder-decoder structure in a U-Net-like fashion, this enables the decoding features of various layers to be constantly concatenated with the corresponding feature of an encoder in the temporal dimension. Also benefit from spatio-temporal modeling of 3D convolution but unlike previous works, a novel saliency decoder is designed in our work to perform aggregation of the multi-scale features.

2.2. Audio-Visual Saliency Prediction

Early audio-visual saliency prediction methods attempted to establish the cross-modal connections between the two modalities by using CCA [32, 33], but an end-to-end deep learning scheme is still far from in-depth study. More recently, Tavakoli *et al*. [37] propose to train two independent 3D ResNet for audio and visual modalities, and the outputs are directly concatenated as a late fusion strategy. With SoundNet [3] for audio representation learning, Tsiami *et al*. [39] perform a spatial sound source localization to obtain audio features, which are then fused with the visual feature maps by bilinear operation. In the same way, Jain *et al*. [26] also employ bilinear fusion operation on the audio features of SoundNet and visual features to predict saliency maps. For these solutions based on a bilinear-based fusion scheme, the large number of learning parameters causes the model learning not easy to converge [38, 48]. In our work, an attention-based fusion is exploited to learn the cross-modal semantic interaction to overcome such a limitation.

2.3. Audio-Visual Consistency Learning

For the intrinsic structure of video stream, the audio is naturally paired and synced with the visual component, which means that the audio-visual correspondence can be effectively utilized to draw direct supervision for different tasks: such as visually guided-source separation [19], audio-visual navigation [11], active speaker detection [46], and audio-visual speech recognition [1], *etc*. Unfortunately, most current audio-visual saliency predictions [26, 39] rely heavily on temporal consistency of visual and audio information, while ignoring the negative impacts of inconsistent

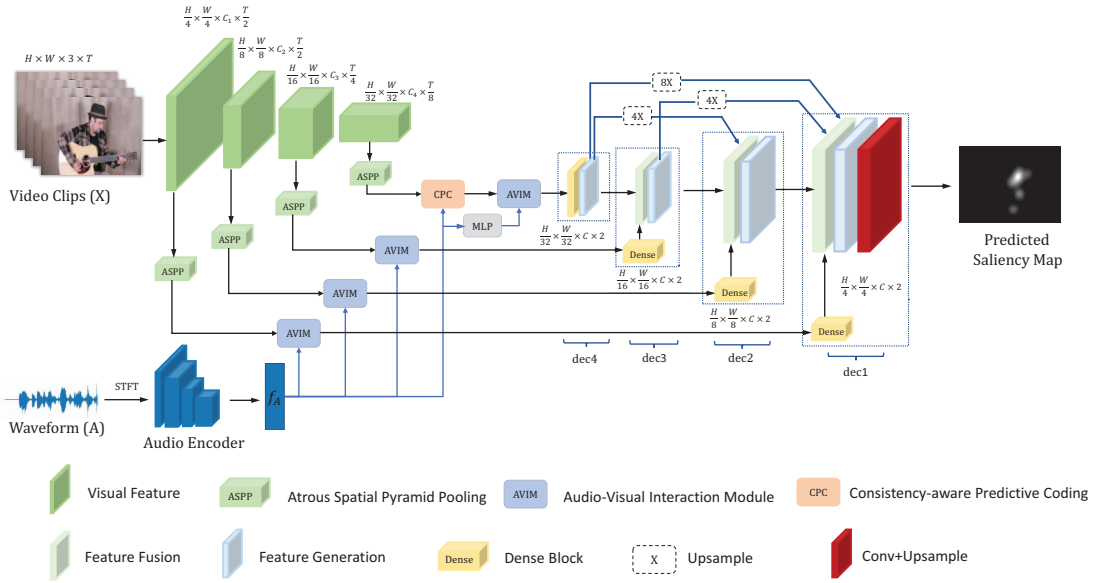

Figure 2. An overview of the proposed CASP-Net. By combining multi-scale visual features and audio features, the designed AVIM and CPC modules enable the network to learn audio-visual semantic interaction and consistent perception. The final saliency decoder utilizes multi-scale audio-visual information to generate saliency maps.

samples. As a promising solution, audio-visual consistency detection can be taken into consideration to ensure the performance of saliency prediction [12,47]. This also becomes a motivation for this study to devise consistency-aware predictive coding that reasons coherent visual and audio content in an iterative feedback manner.

3. CASP-Net

As shown in Figure 2, the proposed CASP-Net is composed of: a two-stream network to obtain visual saliency and auditory saliency feature, an audio-visual interaction module to integrate the visual and auditory conspicuity maps, a consistency-aware predictive coding module to reason the coherent spatio-temporal visual feature with audio feature, and a saliency decoder to estimate saliency map with multi-scale audio-visual features. Each part is elaborated in the below.

3.1. Two-Stream Encoders

Let $X \in \mathbb{R}^{H_v \times W_v \times 3 \times T_v}$ and $A \in \mathbb{R}^{T_A}$ denote video frames and the corresponding audio signal, respectively.

Video Encoder: We employ the off-the-shelf S3D [45] as a video backbone network to encode the spatio-temporal information. This is because that S3D is lightweight and pre-trained on a large dataset, which makes it fast and effective for transfer learning. The backbone consists of 4 convolutional stages, and outputs hierarchical visual feature maps during the encoding process, as shown in Figure 2. The generated features are denoted as $f_{X_i} \in \mathbb{R}^{h_i \times w_i \times C_i \times T_i}$, where $(h_i, w_i) = (H_v, W_v)/2^{i+1}, i = 1, ..., 4$.

Audio Encoder: For audio representation, the 1D audio waveform needs to be converted into the 2D spectrogram by Short-Time Fourier Transform (STFT). Instead of directly applying 1D CNNs on time domain audio signals, a 2D fully convolutional network is employed for this operation. When audio is cropped to match the visual frames duration (*e.g.*, $T_v = 16$), the log-Mel spectrogram is calculated for each matched signal by taking absolute values of a complex STFT, following the natural logarithm. For the high-level semantic information, we employ the VGGish network [23] with pre-trained weights on AudioSet. An audio embedding is generated as the original audio feature $f_A \in \mathbb{R}^{C_A}$ from the layers before the final post-processing stage.

3.2. Cross-Modal Semantic Interaction

To find cross-modal semantics with implicitly associated audio and visual representations in videos, we perform Atrous Spatial Pyramid Pooling (ASPP) [13] on the post-process of visual features f_{X_i} to $f_{V_i} \in \mathbb{R}^{h_i \times w_i \times C \times T_i}$, where $C = 256$. With multiple parallel filters of different rates, the pooling operation helps to recognize visual objects with different receptive fields, *e.g.*, different-sized moving objects.

Considering visual and auditory features have different feature dimensions, an affine transformation is applied to the audio feature to match the channel of visual feature f_{V_i}. Then it is duplicated $h_i w_i T_i$ times in spatio-temporal dimension, and reshaped to the same size as f_{V_i}, which is denoted as $f_{\hat{A}}$. To learn the correspondence between the audio and visual features $f_{\hat{A}}$, f_{V_i}, two different audio-visual interaction approaches are investigated:

Audio-Visual Interaction (Quadratic): For the encoding

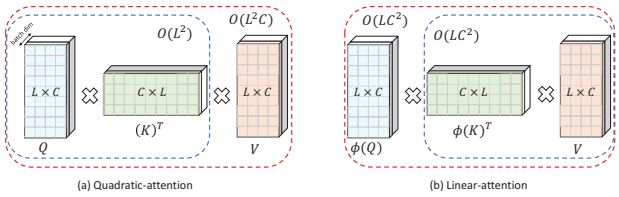

Figure 3. Quadratic vs. Linear. L denotes the number of tokens ($T_i h_i w_i$) in feature, and C denotes a constant (*i.e.*, $C = 256$). Quadratic-attention scales with the *square* of the L. Using a decomposable kernel $\phi(\cdot)$, we can rearrange the order of operations such that linear attention scales *linearly* with L. Dashed-blocks indicate order of computation with corresponding time complexities attached.

of audio-video correlations, the mechanism of self-attention [40] is adopted. According to the calculation method of self-attention , the audio and visual feature matrices need to be transformed into the vector format as $f_{V_i} \in \mathbb{R}^{T_i h_i w_i \times C}$ and $f_{\hat{A}} \in \mathbb{R}^{T_i h_i w_i \times C}$, respectively. Such an audio-visual interaction can be measured by dot-product, then the updated feature maps f_{V_i} at the i-th stage becomes,

$$\begin{aligned} Q &= \alpha(f_{V_i});\ K = \beta(f_{\hat{A}});\ V = \gamma(f_{V_i}) \\ \widetilde{A} &= softmax(\frac{QK^T}{N})V \\ f_{V_i} &= f_{V_i} + \delta(\widetilde{A}) \end{aligned} \tag{1}$$

where α, β, γ and δ are $1 \times 1 \times 1$ convolutions, $N = T_i \times h_i \times w_i$ is a scale factor, and $\widetilde{A}$ denotes the audio-visual similarity matrix. Each visual pixel can be associated with all auditory information by audio-visual interaction.

However, Equation 1 shows that the computational cost of self-attention increases quadratically with the number of tokens ($T_i h_i w_i$) in feature. The same is true for the memory requirements because the similarity matrix $\widetilde{A}$ must be saved to calculate the gradients with respect to the Q, K and V (see also Figure 3(a)).

Audio-Visual Interaction (Linear): As in [27], the audio-visual similarity matrix $\widetilde{A}$ can be generalized by treating $softmax(\cdot)$ as a pairwise similarity between Q and K. That is, for some similarity function $sim(\cdot)$, we have,

$$\widetilde{A} = sim(Q, K)V \tag{2}$$

If choose a decomposable kernel with feature representation $\phi(\cdot) \geq 0$ as $sim(x, y) = \phi(x)\phi(y)^T$, we have

$$\widetilde{A}(\phi) = (\phi(Q)\phi(K)^T)V \tag{3}$$

Then by associativity, the order of computation can be changed as,

$$\widetilde{A}(\phi) = \phi(Q)(\phi(K)^T V) \tag{4}$$

which allows us to compute $\phi(K)^T V$. This leads to an operation $O(T_i h_i w_i C^2)$ to create a C^2 matrix instead of a $(T_i h_i w_i)^2$ one, where C is usually much less than $T_i h_i w_i$ (see also Figure 3(b)). To implement the similarity matrix, the following kernel function is designed as:

$$\phi(x) = gelu(x) + 0.2 \tag{5}$$

where $gelu(\cdot)$ denotes the gaussian error linear units [22].

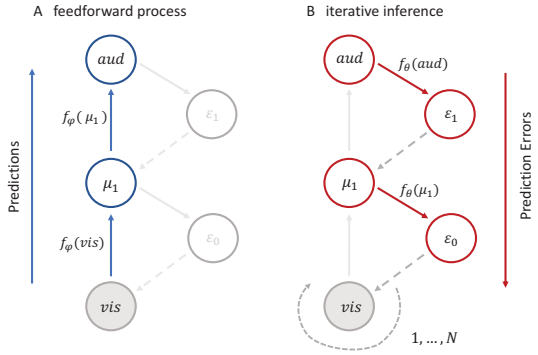

Figure 4. Consistency-aware Predictive Coding combines two different phases as follows. For brevity, we use vis and aud to denote f_{V_4} and f_A, respectively. (A) The vis propagated up the hierarchy in a feedforward manner, utilising the non-linear function $f_\varphi(\cdot)$. (B) The initial values for μ are then used to predict the activity at the layer below, transformed by the iterative functions $f_\theta(\cdot)$. These predictions incur prediction error ϵ, which are then used to update activity μ. This process is repeated N times, after which perceptual inference is complete.

3.3. Consistency-aware Predictive Coding

To overcome the potential inconsistencies introduced by audio and visual features, consistency-aware predictive coding (CPC) is proposed to improve the performance of reasoning multi-modal features. Inspired by the predictive coding (PC) in theoretical neuroscience [2, 6], it represents top-down signalling in the perceptual hierarchy to predict the cause of sensory data. For predictive coding, each hierarchical layer predicts the activity of the layer below it (with the lowest layer predicting the sensory data). Such predictions are then iteratively refined by minimizing the prediction errors, (*i.e.*, the difference between predictions and the actual activity), in each layer [18, 36]. Comparatively, the proposed CPC adopts bottom-up $f_\varphi(\cdot)$ and top-down $f_\theta(\cdot)$ paths to represent the prediction and iterative process respectively, as shown in Figure 4. The CPC inputs the visual feature vis as a type of prior knowledge to predict the audio feature aud iteratively. It is composed of L hierarchical layers, in which each layer i is composed of a variable unit μ_i and an error unit ϵ_i.

Feedforward Process: A set of feedforward parameters φ are defined in correspondence to bottom-up connections. The feedforward parameters represent non-linear functions which map activity at one layer to activity at above layer. Thereby a bottom-up prediction is implemented as:

$$\mu_i = f_\varphi(\mu_{i-1}) \tag{6}$$

Iterative Inference: The iterative phase updates the activity μ_i again by generating top-down prediction $f_\theta(\mu_{i+1})$ with the prediction error ϵ_i^{pred}.

$$\begin{aligned} \epsilon_i^{pred} &= MSE(\mu_i - f_\theta(\mu_{i+1})) \\ \mu_i &\leftarrow \mu_i - \alpha \nabla \epsilon_i^{pred} \end{aligned} \tag{7}$$

where $\nabla \epsilon_i^{pred}$ denotes the gradient with respect to the activity of μ_{i+1}, the hyper-parameter α is introduced to reduce the prediction error.

To decrease the prediction error in CPC, the operations of feedforward process and iterative inference are alternately performed while gradually improving the representations of all layers. When the entire inference ends, the bottom layer is directly output with perceptually consistent audio-visual features.

3.4. SalDecoder

A new decoder architecture is also proposed for saliency estimation. The decoder consists of 4 blocks, *i.e.*, dec_x, where $x = 1, ..., 4$. Figure 2 presents the overview of the decoder's block. Each of the blocks consists of a dense block [25] and a feature generation block. Moreover, the blocks dec_1, dec_2, and dec_3 have a fusion function among them named fusion blocks, which combine the output of the previous decoder with the output from the dense block. The final saliency map is obtained using the combination of Conv and Upsample in the dec_1 stage.

Specifically, the audio-visual feature representations $\{f_{X_i}\}(i = 1, ..., 4)$ are taken as input from AVIM to the decoder. In each stream, the dense block is firstly utilized to process representation by its feature propagation, and align the temporal dimensions of each feature to facilitate subsequent fusion. All blocks except dec_4 contain a fusion block to integrate multi-scale features. For the explicit operation in the fusion block, each feature is initialized to align in the spatial dimension via upsampling, and fused by element-wise summation. The fused feature are fed into the feature generation block, which consists of BN-ReLU-3D Conv layers, to obtain the semantic features with context information. In the dec_1, the combination of 3D Conv and Upsample is also performed to generate the final saliency map.

3.5. Saliency Losses

We refer to the training paradigm of multiple loss functions in [10, 39], which contains: Kullback-Leibler (KL) divergence, Linear Correlation Coefficient (CC) and Similarity Metric (SIM). Assuming that the predicted saliency map is $S_{pred} \in [0, 1]$, the labeled binary fixation map is $S_{fix} \in \{0, 1\}$, and the dense saliency map generated by the fixation map is $S_{den} \in [0, 1]$, then L_{KL}, L_{CC}, and L_{NSS} are employed to signify three different loss functions, respectively. The first is the KL loss between the predicted map S_{pred} and the dense map S_{den}:

Data	Method	#Params	CC↑	NSS↑	SIM↑
AVA	CASP-Net(FCN)	**1.77**	0.659	3.57	0.509
	CASP-Net(UNet)	3.94	0.665	3.65	0.512
	CASP-Net(Sal)	2.49	**0.671**	**3.67**	**0.515**
ETMD	CASP-Net(FCN)	**1.77**	0.606	3.21	0.463
	CASP-Net(UNet)	3.94	0.611	3.27	0.468
	CASP-Net(Sal)	2.49	**0.613**	**3.30**	**0.471**

Table 1. Comparison between SalDecoder and the other different decoders on AVAD and ETMD datasets (visual-only). #Params represent the number of parameters per decoder.

$$L_{KL}(S_{pred}, S_{den}) = \sum_x S_{den}(x) ln \frac{S_{den}(x)}{S_{pred}(x)} \tag{8}$$

where x represents the spatial domain of a saliency map. The second loss function is based on the CC that has been widely used in saliency evaluation, and used to measure the linear relationship between the predicted saliency map S_{pred} and the dense map S_{den}:

$$L_{CC}(S_{pred}, S_{den}) = -\frac{cov(S_{pred}, S_{den})}{\rho(S_{pred})\rho(S_{den})} \tag{9}$$

where $cov(\cdot)$ and $\rho(\cdot)$ represent the covariance and the standard deviation respectively. The last one is derived from the SIM, which can measure the similarity between two distributions:

$$L_{SIM}(S_{pred}, S_{den}) = \sum_x min\{\zeta(S_{pred}(x)), \zeta(S_{den}(x))\} \tag{10}$$

where ζ represents the normalization operation. The weighted summation of the above KL, CC and SIM is taken to represent the final loss function:

$$L_{total} = L_{KL} + \lambda_1 L_{CC} + \lambda_2 L_{SIM} \tag{11}$$

where λ_1, λ_2 are the weights of CC and SIM, respectively.

4. Experiment

Experiments are conducted on a total of seven datasets including a pure visual dataset and six audio-visual eye-tracking datasets. In the following subsections, the implementation details and evaluation metrics are firstly introduced. We represent the experimental results with analysis via the ablation studies and comparison with the state-of-the-art works.

4.1. Setup

4.1.1 Datasets

Visual Dataset: The DHF1k [43] is one of the most popular visual-only datasets in the study of video saliency, It contains 1000 videos where 600 videos are for training and 100 for validation. In addition, a test set of 300 videos is also released but without public ground truth. Considering that the main focus of our model is on multi-modal scenarios, its visual branch is pre-trained using this dataset.

Method	AVAD		ETMD	
	CC ↑	SIM ↑	CC ↑	SIM ↑
Visual-Only	0.671	0.515	0.613	0.469
V+A+Bilinear	0.670	0.510	0.609	0.469
V+A+I$_{Quadratic}$	0.674	0.517	0.615	0.471
V+A+I$_{Linear}$	0.675	0.519	0.615	0.470
V+A+I$_{Linear}$ +C	**0.685**	**0.528**	**0.616**	**0.476**

Table 2. Ablation Studies. The visual-only (V) denotes the visual branch of CASP-Net. A refers to the audio branch, I refers to the audio-visual interaction module, C refers to the consistency-aware predictive coding, and the subscripts represent two schemes with different computational complexity.

Data	Metric	i-th stage of Video Encoder, $i \in \{1,2,3,4\}$					
		1	2	3	4	2,3,4	1,2,3,4
AVA	CC	0.670	0.672	**0.673**	0.671	0.674	**0.675**
	SIM	0.510	0.512	**0.514**	0.513	0.515	**0.518**
ETMD	CC	0.610	**0.613**	0.611	0.611	0.613	**0.615**
	SIM	0.467	**0.469**	0.468	0.467	**0.471**	0.470

Table 3. Audio-visual Interaction at various video encoder stages. In both the AVAD and ETMD datasets, the model achieves almost the best performance when the AVIM is used in all four stages.

Audio-Visual Dataset: There are six audio-visual datasets in video saliency: AVAD [32], Coutrot1 [16], Coutrot2 [17] , DIEM [34], ETMD [28], and SumMe [21], which are used for our evaluation comparison. In detail, (i) the AVAD dataset contains 45 video clips with a duration of 5-10 seconds. These clips cover a variety of audio-visual activities, *e.g.*, playing the piano, playing basketball, making an interview, *etc*. The dataset contains eye-tracking data from 16 participants. (ii) The Coutrot1 and Coutrot2 datasets are separated from the Coutrot dataset. The Coutrot1 dataset contains 60 video clips covering 4 visual categories: one moving object, several moving objects, landscapes, and faces. The corresponding eye-tracking data are from 72 participants. The Coutrot2 dataset contains 15 video clips, which record 4 persons having a meeting. The corresponding eye-tracking data are from 40 participants. (iii) The DIEM dataset contains 84 video clips including game trailers, music videos, advertisements and *etc*. which are captured from 42 participants. It should be noted that the audio and visual tracks in these videos do not correspond naturally. The ETMD dataset contains 12 video clips from several Hollywood movies, with the eye-tracking data annotated by 10 different people. The SumMe dataset consists of 25 video clips with diverse topics, *e.g.*, playing ball, cooking, travelling, *etc*., and the corresponding eye-tracking data are collected from 10 viewers.

4.1.2 Implementation Details

We use pre-trained S3D model [45] on Kinetics [9] and pre-trained VGGish [23] on AudioSet [20]. The input samples of the network consist of 16-frame video clips of size $224 \times 384 \times 3$ with the corresponding audio stream, which is transformed into 96×64 log-Mel spectrograms. For a fair comparison, two different training strategies are designed depending on whether the DHF1k dataset is used or not. The first is to train the entire model on the six audio-visual datasets from scratch. The other is to train the visual branch of the model on the DHF1k dataset, and then use this weight to fine-tune the entire model on these audio-visual datasets. Both strategies end up with the same evaluation as [39].

For the CPC module, we mainly adopt 3D Conv layers in the feedforward process and 3D Deconv layers in the iterative inference. The hyper-parameter α is set to 0.1. The proposed training process chooses Adam as the optimizer with the started learning rate of 1e-4. The loss weight is to -0.1, *i.e.*, $\lambda_1 = \lambda_2 = -0.1$. The computation platform is configured by an NVIDIA GeForce RTX 3090 GPU with batch-size 8 for entire experiments.

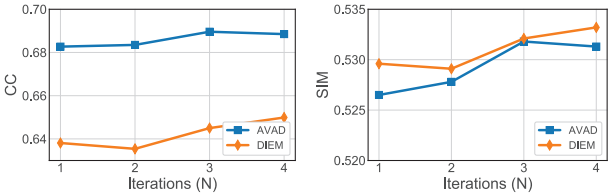

Figure 5. Performance analysis of CPC's iterations on AVAD and DIEM datasets.

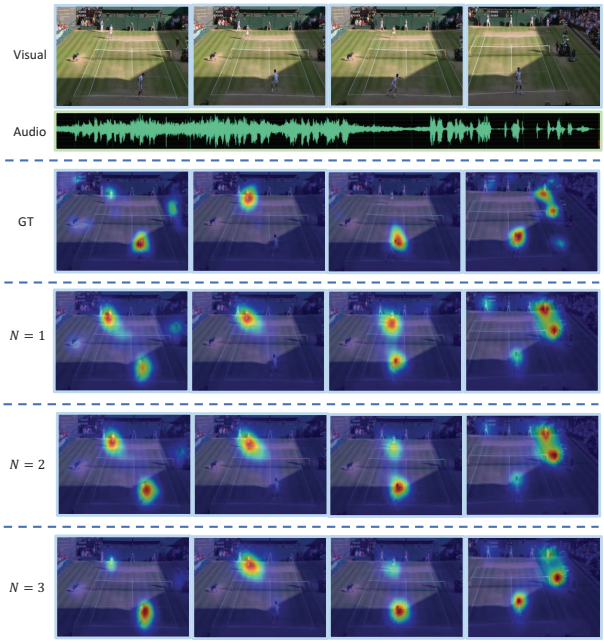

Figure 6. Comparison of saliency maps with different CPC's iterations. *GT* denotes ground truth.

4.1.3 Evaluation metrics

For the evaluation of CASP-Net, four widely-used evaluation metrics are adopted [8]: CC, NSS, AUC-Judd (AUC-

Method	Pretrained	DIEM				Coutrot1				Coutrot2			
		CC ↑	NSS ↑	AUC-J ↑	SIM ↑	CC ↑	NSS ↑	AUC-J↑	SIM ↑	CC↑	NSS↑	AUC-J↑	SIM↑
ACLNet(V) [44]	-	0.522	2.02	0.869	0.427	0.425	1.92	0.850	0.361	0.448	3.16	0.926	0.322
TASED-Net(V) [31]	-	0.557	2.16	0.881	0.461	0.479	2.18	0.867	0.388	0.437	3.17	0.921	0.314
STAViS(V) [39]	-	0.567	2.19	0.879	0.472	0.458	1.99	0.861	0.384	0.652	4.19	0.940	0.447
STAViS(AV) [39]	-	0.579	2.26	0.883	0.482	0.472	2.11	0.868	0.393	0.734	5.28	0.958	0.511
CASP-Net(V)	-	0.638	2.54	0.902	0.529	0.555	2.65	0.882	0.449	0.756	6.01	0.961	0.566
CASP-Net(AV)	-	**0.649**	**2.58**	**0.904**	**0.536**	**0.560**	**2.66**	**0.887**	**0.453**	**0.766**	**6.11**	**0.963**	**0.573**
ViNet(V) [26]	DHF1k	0.626	2.47	0.898	0.483	0.551	2.68	0.886	0.423	0.724	5.61	0.95	0.466
ViNet(AV) [26]	DHF1k	0.632	2.53	0.899	0.498	0.56	2.73	0.889	0.425	0.754	5.95	0.951	0.493
TSFP-Net(V) [10]	DHF1k	0.649	**2.63**	0.905	0.529	0.57	**2.75**	0.894	0.451	0.718	5.30	0.957	0.516
TSFP-Net(AV) [10]	DHF1k	0.651	2.62	0.906	0.527	**0.571**	2.73	**0.895**	0.447	0.743	5.31	0.959	0.528
CASP-Net(V)	DHF1k	0.649	2.59	0.904	0.538	0.559	2.64	0.888	0.445	0.756	6.07	0.963	0.567
CASP-Net(AV)	DHF1k	**0.655**	2.61	**0.906**	**0.543**	0.561	2.65	0.889	**0.456**	**0.788**	**6.34**	**0.963**	**0.585**

Table 4. Comparison of saliency on DIEM, Coutrot1 and Coutrot2 datasets. The experimental table is divided into two groups according to whether the DHF1k dataset is used as pre-training data. We show the modalities used for each method in brackets: (V) for visual, and (AV) for audio-visual.

Method	Pretrained	AVAD				ETMD				SumMe			
		CC ↑	NSS ↑	AUC-J ↑	SIM ↑	CC↑	NSS↑	AUC-J↑	SIM↑	CC↑	NSS↑	AUC-J ↑	SIM↑
ACLNet(V) [44]	-	0.580	3.17	0.905	0.446	0.477	2.36	0.915	0.329	0.379	1.79	0.868	0.296
TASED-Net(V) [31]	-	0.601	3.16	0.914	0.439	0.509	2.63	0.916	0.366	0.428	2.1	0.884	0.333
STAViS(V) [39]	-	0.604	3.07	0.915	0.443	0.560	2.84	0.929	0.412	0.418	1.98	0.884	0.332
STAViS(AV) [39]	-	0.608	3.18	0.919	0.457	0.569	2.94	0.931	0.425	0.422	2.04	0.888	0.337
CASP-Net(V)	-	0.671	3.67	0.931	0.515	0.613	3.30	0.938	0.471	0.481	2.50	0.901	0.374
CASP-Net(AV)	-	**0.685**	**3.77**	**0.932**	**0.528**	**0.616**	**3.31**	**0.939**	**0.476**	**0.486**	**2.52**	**0.904**	**0.377**
ViNet(V) [26]	DHF1k	0.694	**3.82**	0.928	0.504	0.569	3.06	0.928	0.409	0.466	2.40	0.898	0.345
ViNet(AV) [26]	DHF1k	0.674	3.77	0.927	0.491	0.571	3.08	0.928	0.406	0.463	2.41	0.897	0.343
TSFP-Net(V) [10]	DHF1k	0.688	3.79	0.931	**0.530**	0.576	3.09	0.932	0.433	0.463	2.28	0.894	0.362
TSFP-Net(AV) [10]	DHF1k	**0.704**	3.77	0.932	0.521	0.576	3.07	0.932	0.428	0.464	2.30	0.894	0.360
CASP-Net(V)	DHF1k	0.681	3.75	0.931	0.526	0.616	3.31	0.938	0.471	0.485	2.52	0.904	0.382
CASP-Net(AV)	DHF1k	0.691	3.81	**0.933**	0.528	**0.620**	**3.34**	**0.940**	**0.478**	**0.499**	**2.60**	**0.907**	**0.387**

Table 5. Comparison of saliency on AVAD, ETMD and SumMe datasets. The experimental table is divided into two groups according to whether the DHF1k dataset is used as pre-training data. We show the modalities used for each method in brackets: (V) for visual, and (AV) for audio-visual.

J), and SIM. The CC measures the linear correlation coefficient between the ground truth and the predicted saliency map. The NSS focuses on measuring the saliency value on human fixations, the AUC-J is a location-based metric for evaluating the predicted saliency map, and the SIM measures the similarity between the predicted saliency map and the ground truth.

4.2. Ablation Studies

Table 1 and Table 2 show ablation studies on different configurations of CASP-Net. All the ablations are performed with the training on the AVAD and ETMD training sets and evaluated on their validation sets. From Table 1, three different architecture of decoders are compared w.r.t. their parameters and performance, namely FCN, UNet and Sal (ours). The FCN decoder simply uses multi-layers network with architecture: 3D Deconv + ReLU + BN. UNet decoder represents the progressive upsampling process that alternates 3D Conv and Trilinear layers while adding UNet skip connections operation. Experimental results show that CASP-Net(Sal) outperforms CASP-Net(FCN) and CASP-Net(UNet) with a smaller parameter quantity. It suggests that the decoder FCN and UNet can be replaced in tasks of dense prediction such as saliency prediction, and further confirms the effectiveness of our proposed decoder.

To analyze the performance of each part in the proposed work, a baseline model *Visual-Only* is formed based on a visual encoder and a SalDecoder as shown in Table 2. Moreover, the Bilinear fusion operation [38] is also introduced to compare with the designed AVIM, the obtained observations are listed below: (i) The combination of AVIM and CPC can continuously improve the performance of the model, which achieves the best performance, (ii) The linear version of AVIM has roughly the same performance as the quadratic version, but the computational complexity is much lower, (iii) The Bilinear-based model performs much worse than the AVIM-based model, and even not better than the *Visual-Only* model. This means that the ability of audio-visual feature aggregation from Bilinear operation is inferior to that of the proposed AVIM.

Cross-modal Interaction at Various Stages. For cross-modal interaction, the AVIM has a plug-in architecture that can be applied in any stage. As shown in the Table 3, the prediction performance fluctuates when the AVIM is used in different single stage. On both AVAD and ETMD datasets, it is noticed that AVIM performs better in the second and

third stages because the acquired semantics of the visual features is limited in the initial stage. Since our SalDecoder adopts a skip-connection, it would be beneficial to apply the AVIM in multiple stages, as verified in the right part of Table 3. On both datasets, the best performance is achieved by applying AVIM at four stages, which also indicates the model has the ability to fuse and balance the features from multiple stages.

Further Analysis of CPC. The impact of the iterative number of the CPC on its performance also needs to be analyzed. Figure 5 shows that the prediction metrics tend to rise when more iterative computations are performed, especially in the first three iterations. Thus, the iteration number N is set to 3 by default. Figure 6 depicts the predicted saliency maps of our method on some audio-visual samples, which shows that the CPC iteratively resolves the internal inconsistencies in audio and visual features. It is also noticed that the saliency maps are inaccurate in early iterations, till later iterations, the model has corrected itself to pay more accurate attention to the objects of interest.

4.3. Comparisons with State-of-the-art Methods

The proposed CASP-Net is compared with recent state-of-the-art saliency works on six audio-visual datasets as shown in Table 4 and Table 5. The experiment results can be divided into two groups according to whether the DHF1k dataset is used as pre-training data. Results in the two groups highlight the superiority of the proposed audio-visual scheme, as it outperforms the other state-of-the-art works on almost all datasets and metrics. It can be observed that the CASP-Net is able to significantly surpass the prior saliency predictions, such as STAViS [39] and TASED-Net [31], with the configuration trained on six audio-visual datasets directly. The CASP-Net(AV) achieves an average performance improvement of 11.5% CC and 13% SIM compared to STAViS(AV), and becomes a new state-of-the-art on six benchmarks.

The proposed work also shows an obvious superiority in audio-visual saliency prediction with the DHF1k pre-training dataset. Compared to ViNet [26] and TSFP-Net [10], the CASP-Net(AV) achieves the best performance in most datasets, especially on the Countrot2, ETMD and SumMe test sets. Moreover, after incorporating audio features into the pure visual model, the relative improvement on these three datasets is the most significant (average 3 % relative improvement in CC and 2 % relative improvement in SIM). This reflects the high audio-visual correspondence of three datasets, as well as the model's capability to take full advantage of these audio-visual cues. Overall, the experiment results in both groups have shown that pre-training on a large-scale video dataset enables the proposed CASP-Net to be effectively generalized on the other datasets. For qualitative analysis, the CASP-Net is further compared with the STAViS and ViNet on Coutrot2, ETMD, AVAD and SumMe datasets to show superior performance.

In Figure 7, the advantages of the proposed work have been demonstrated in the audio-visual scenarios having multiple speakers. The first and second rows are the frames and the corresponding ground truth saliency map. The third row presents the prediction saliency maps from our CASP-Net, and the final 2 rows are the same maps for STAViS and ViNet. In particular, our results are closer to the ground truth.

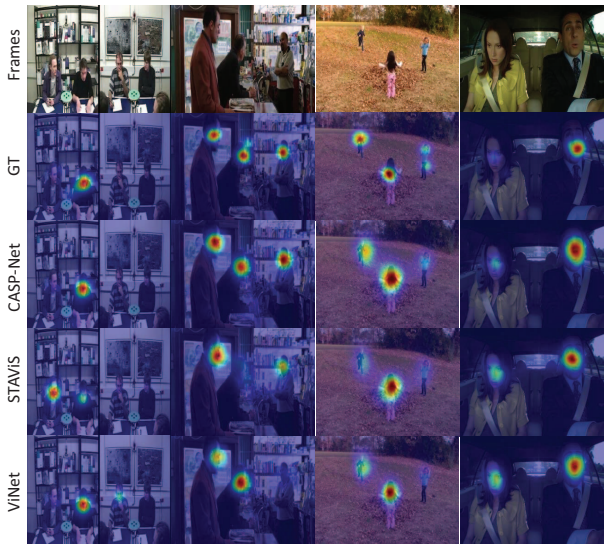

Figure 7. Sample frame from Coutrot2, ETMD, AVAD and SumMe databases with their eye-tracking data, and the corresponding ground truth, CASP-Net, and other state-of-the-art audio-visual saliency maps for comparisons.

5. Conclusion

We propose a consistency-aware audio-visual saliency prediction network (CASP-Net), which effectively addresses potential audio-visual inconsistency in video saliency prediction. A two-stream network is designed to elegantly associate video frames with the corresponding sound source, achieving cross-modal semantic similarities between audio and visual features. In addition, a novel consistency-aware predictive coding module is introduced to improve the consistency within audio and visual representations iteratively. Besides, a saliency decoder is also designed to aggregate the multi-scale audio-visual information and obtain the final saliency map. Experiments show surprising results that CASP-Net outperforms 5 state-of-the-art approaches on 6 challenging audio-visual eye-tracking datasets.

Acknowledgements. This work was supported by NSFC under Grants 61971352, 62271239, 61862043, Ningbo Natural Science Foundation under Grants 2021J048, 2021J049.

References

[1] T. Afouras, J. S. Chung, and A. Zisserman. Deep lip reading: a comparison of models and an online application. In *INTERSPEECH*, 2018. 2

[2] Arjen Alink, Caspar M Schwiedrzik, Axel Kohler, Wolf Singer, and Lars Muckli. Stimulus predictability reduces responses in primary visual cortex. *Journal of Neuroscience*, 30(8):2960–2966, 2010. 4

[3] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. *Advances in neural information processing systems*, 29, 2016. 2

[4] Cagdas Bak, AYSUN KOÇAK, MEHMET ERDEM, and İBRAHİM Erdem. Spatio-temporal saliency networks for dynamic saliency prediction. *IEEE TRANSACTIONS ON MULTIMEDIA*, 20(7), 2018. 1, 2

[5] Giovanni Bellitto, Federica Proietto Salanitri, Simone Palazzo, Francesco Rundo, Daniela Giordano, and Concetto Spampinato. Video saliency detection with domain adaption using hierarchical gradient reversal layers. *arXiv preprint arXiv:2010.01220*, 2020. 2

[6] Christopher L Buckley, Chang Sub Kim, Simon McGregor, and Anil K Seth. The free energy principle for action and perception: A mathematical review. *Journal of Mathematical Psychology*, 81:55–79, 2017. 4

[7] Nicholas J Butko, Lingyun Zhang, Garrison W Cottrell, and Javier R Movellan. Visual saliency model for robot cameras. In *2008 IEEE International Conference on Robotics and Automation*, pages 2398–2403. IEEE, 2008. 1

[8] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. What do different evaluation metrics tell us about saliency models? *IEEE transactions on pattern analysis and machine intelligence*, 41(3):740–757, 2018. 6

[9] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 6

[10] Qinyao Chang and Shiping Zhu. Temporal-spatial feature pyramid for video saliency detection. *arXiv preprint arXiv:2105.04213*, 2021. 5, 7, 8

[11] Changan Chen, Ziad Al-Halah, and Kristen Grauman. Semantic audio-visual navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15516–15525, 2021. 2

[12] Chenglizhao Chen, Mengke Song, Wenfeng Song, Li Guo, and Muwei Jian. A comprehensive survey on video saliency detection with auditory information: the audio-visual consistency perceptual is the key! *IEEE Transactions on Circuits and Systems for Video Technology*, 2022. 3

[13] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 3

[14] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014. 2

[15] Runmin Cong, Jianjun Lei, Huazhu Fu, Ming-Ming Cheng, Weisi Lin, and Qingming Huang. Review of visual saliency detection with comprehensive information. *IEEE Transactions on circuits and Systems for Video Technology*, 29(10):2941–2959, 2018. 1

[16] Antoine Coutrot and Nathalie Guyader. How saliency, faces, and sound influence gaze in dynamic social scenes. *Journal of vision*, 14(8):5–5, 2014. 6

[17] Antoine Coutrot and Nathalie Guyader. Multimodal saliency models for videos. In *From Human Attention to Computational Attention*, pages 291–304. Springer, 2016. 6

[18] Karl Friston. A theory of cortical responses. *Philosophical transactions of the Royal Society B: Biological sciences*, 360(1456):815–836, 2005. 2, 4

[19] Ruohan Gao and Kristen Grauman. Visualvoice: Audio-visual speech separation with cross-modal consistency. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15490–15500. IEEE, 2021. 2

[20] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017. 6

[21] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, and Luc Van Gool. Creating summaries from user videos. In *European conference on computer vision*, pages 505–520. Springer, 2014. 6

[22] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 4

[23] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE, 2017. 3, 6

[24] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 2

[25] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 5

[26] Samyak Jain, Pradeep Yarlagadda, Shreyank Jyoti, Shyamgopal Karthik, Ramanathan Subramanian, and Vineet Gandhi. Vinet: Pushing the limits of visual modality for audio-visual saliency prediction. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3520–3527. IEEE, 2021. 2, 7, 8

[27] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International Conference on Machine Learning*, pages 5156–5165. PMLR, 2020. 4

[28] Petros Koutras and Petros Maragos. A perceptually based spatio-temporal computational framework for visual saliency

estimation. *Signal Processing: Image Communication*, 38:15–31, 2015. 6

[29] Qiuxia Lai, Wenguan Wang, Hanqiu Sun, and Jianbing Shen. Video saliency prediction using spatiotemporal residual attentive networks. *IEEE Transactions on Image Processing*, 29:1113–1126, 2019. 2

[30] Viraj Mavani, Shanmuganathan Raman, and Krishna P Miyapuram. Facial expression recognition using visual saliency and deep learning. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 2783–2788, 2017. 1

[31] Kyle Min and Jason J Corso. Tased-net: Temporally-aggregating spatial encoder-decoder network for video saliency detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2394–2403, 2019. 2, 7, 8

[32] Xiongkuo Min, Guangtao Zhai, Ke Gu, and Xiaokang Yang. Fixation prediction through multimodal analysis. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 13(1):1–23, 2016. 2, 6

[33] Xiongkuo Min, Guangtao Zhai, Jiantao Zhou, Xiao-Ping Zhang, Xiaokang Yang, and Xinping Guan. A multimodal saliency model for videos with high audio-visual correspondence. *IEEE Transactions on Image Processing*, 29:3805–3819, 2020. 1, 2

[34] Parag K Mital, Tim J Smith, Robin L Hill, and John M Henderson. Clustering of gaze during dynamic scene viewing is predicted by motion. *Cognitive computation*, 3(1):5–24, 2011. 6

[35] Tam V Nguyen, Mengdi Xu, Guangyu Gao, Mohan Kankanhalli, Qi Tian, and Shuicheng Yan. Static saliency vs. dynamic saliency: a comparative study. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 987–996, 2013. 1

[36] Rajesh PN Rao and Dana H Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, 2(1):79–87, 1999. 2, 4

[37] Hamed R Tavakoli, Ali Borji, Esa Rahtu, and Juho Kannala. Dave: A deep audio-visual embedding for dynamic saliency prediction. *arXiv preprint arXiv:1905.10693*, 2019. 1, 2

[38] Joshua B Tenenbaum and William T Freeman. Separating style and content with bilinear models. *Neural computation*, 12(6):1247–1283, 2000. 2, 7

[39] Antigoni Tsiami, Petros Koutras, and Petros Maragos. Stavis: Spatio-temporal audiovisual saliency network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4766–4776, 2020. 1, 2, 5, 6, 7, 8

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4

[41] Eleonora Vig, Michael Dorr, and David Cox. Large-scale optimization of hierarchical features for saliency prediction in natural images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2798–2805, 2014. 1

[42] Wenguan Wang and Jianbing Shen. Deep visual attention prediction. *IEEE Transactions on Image Processing*, 27(5):2368–2378, 2017. 1

[43] Wenguan Wang, Jianbing Shen, Fang Guo, Ming-Ming Cheng, and Ali Borji. Revisiting video saliency: A large-scale benchmark and a new model. In *Proceedings of the IEEE Conference on computer vision and pattern recognition*, pages 4894–4903, 2018. 1, 2, 5

[44] Wenguan Wang, Jianbing Shen, Jianwen Xie, Ming-Ming Cheng, Haibin Ling, and Ali Borji. Revisiting video saliency prediction in the deep learning era. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):220–237, 2019. 7

[45] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European conference on computer vision (ECCV)*, pages 305–321, 2018. 2, 3, 6

[46] Junwen Xiong, Yu Zhou, Peng Zhang, Lei Xie, Wei Huang, and Yufei Zha. Look&listen: Multi-modal correlation learning for active speaker detection and speech enhancement. *IEEE Transactions on Multimedia*, 2022. 2

[47] Hanyu Xuan, Zhenyu Zhang, Shuo Chen, Jian Yang, and Yan Yan. Cross-modal attention network for temporal inconsistent audio-visual event localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 279–286, 2020. 3

[48] Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 1821–1830, 2017. 2

[49] Kao Zhang, Zhenzhong Chen, and Shan Liu. A spatial-temporal recurrent neural network for video saliency prediction. *IEEE Transactions on Image Processing*, 30:572–587, 2020. 1

[50] Mengmeng Zhu, Guanqun Hou, Xinjia Chen, Jiaxing Xie, Haixian Lu, and Jun 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*, pages 1953–1962, 2021. 1

[51] Shiping Zhu and Ziyao Xu. Spatiotemporal visual saliency guided perceptual high efficiency video coding with neural network. *Neurocomputing*, 275:511–522, 2018. 1