Does Audio help in deep Audio-Visual Saliency prediction models?

Thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Science in
Electronics and Communication Engineering
by Research

by

Ritvik Agrawal
2018122005
ritvik.agrawal@research.iiit.ac.in

International Institute of Information Technology
Hyderabad - 500 032, INDIA
April 2023

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled **" Does Audio help in deep Audio-Visual Saliency prediction models?"** by **Ritvik Agrawal**, has been carried out under my supervision and is not submitted elsewhere for a degree.

_____ _____

Date Adviser: Dr. Vineet Gandhi

To my Family

Acknowledgments

Completion of this dissertation is possible with the support of many people. I would like to express my sincere gratitude to all of them.

Foremost, I would like to express my sincere gratitude to my advisor, *Dr. Vineet Gandhi*, for his patience, motivation, immense knowledge, and invaluable guidance. I am deeply obliged to him for being my mentor and his constant cooperation and support has always motivated me to keep going ahead. I gratefully acknowledge his constant efforts to instill good research skills, values, and ethics in me.

I would like to thank my CVIT lab mates and coauthors: Rohit Girmaji, Sarath Sivaprasad, and Shreyank Jyoti (names in alphabetical order) for constantly helping and guiding me, from explaining numerous concepts to helping me with code implementation.

I also thank the overall CVIT ecosystem for the unparalleled access to resources. In particular, the institute cluster "Ada", named in honor of Ada Lovelace has been a vital aspect in all the work done as part of this thesis.

A debt of gratitude is owed to my Family, for their unconditional love and moral support at every stage of my life. All my accomplishments are incomplete and impossible without them.

Above all, I owe it all to Almighty God for granting me wisdom, health, and strength to undertake this research task and enabling me to its completion.

Abstract

The task of saliency prediction focuses on understanding and modeling human visual attention (HVA), i.e., where and what people pay attention to given visual stimuli. Audio has ideal properties to aid gaze fixation while viewing a scene. There exists substantial evidence of audio-visual interplay in human perception, and it is agreed that they jointly guide our visual attention. Learning computational models for saliency estimation is an effort to inch machines/robots closer to human cognitive abilities. The task of saliency prediction is helpful in many digital content-based applications like automated editing, perceptual video coding, human-robot interactions,etc.

The field has progressed from using hand-crafted features to deep learning-based solutions. Efforts on static image saliency prediction methods are led by convolutional architectures. The ideas were extended to videos by integrating temporal information using 3D convolutions or LSTM's. Many sophisticated multimodal, multi-stream architectures have been proposed to process multimodal information for saliency prediction.

Despite existing works of Audio-Visual Saliency Prediction (AVSP) models claiming to achieve promising results by fusing audio modality over visual-only models, most of these models only consider visual cues and fail to leverage auditory information that is ubiquitous in dynamic scenes. In this thesis, we investigate the relevance of audio cues in conjunction with the visual ones and conduct extensive analysis to analyse the cause of AVSP models being superior by employing well-established audio modules and fusion techniques from diverse correlated audio-visual tasks. Our analysis on ten diverse saliency datasets suggests that none of the methods worked for incorporating audio. Our endeavour suggests that augmenting audio features ends up learning a predictive model agnostic to audio . Furthermore, we bring to light, why AVSP models show a gain in performance over visual-only models, though the audio branch is agnostic at inference.

Our experiments clearly indicate that visual modality dominates the learning; the current models largely ignore the audio information. The observation is consistent while using three different audio backbones and four different fusion techniques and contrasts with the previous methods, which claim audio as a significant contributing factor. The performance gains are a byproduct of improved training and the additional audio branch seems to have a regularizing effect. We show that similar gains are achieved while sending random audio during training. Overall our work questions the role of audio in current deep AVSP models and motivates the community to a clear avenue for reconsideration of the complex architectures by demonstrating that simpler alternatives work equally well.

Contents

List of Figures

List of Tables

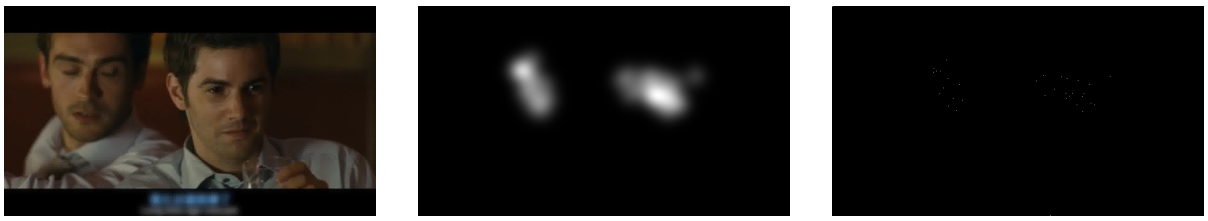
Chapter 1

Introduction

Figure 1.1: Example frames from a multi-person conversation video (first column) and the corresponding saliency (second column). ViNet is a visual-only saliency prediction model, and AViNet is an audio-visual saliency prediction model. AViNet gives better predictions in this example. On the first reflection, it appears that audio plays a key role. However, when performing inference with zero audio *[AViNet-0]* and random audio *[AViNet-R]*, the output predictions are identical. Clearly, the audio is obsolete at inference. Our work finds that the audio branch may merely act as a regularizer and motivates a review of multi-modal interaction in audio-visual saliency prediction models.

1.1 Motivation and Background

The human visual attention (HVA) mechanism facilitates diverse information processing in our surroundings by localizing the most prominent (salient) region[26]. Predicting the salient regions in a scene is a fundamental ability, which empowers primates to rapidly analyze/interpret the complex surroundings by locating and devoting the focus only on sub-regions of interest [30]. Mimicking this ability in machines is a fundamental research problem [5] and is actively pursued in the domains of computer vision, cognitive science, robotics, and human-computer interaction. A primary way to address the problem is to first compile ground truth regarding where viewers gaze in the scene via eye-tracking hardware, train machine learning models, and perform prediction on novel unseen video computationally. This task is commonly referred as saliency prediction and is shown to be effective in many downstream applications such as video surveillance [72], cinematic editing [47], video captioning [49], virtual reality [57], video compression[25], human-robot interaction [20], scene classification[4], region tracking [21], proposal refinement[10], etc., owing to it's ability to prioritize the video information across space and time.



(a) RGB Image (b) Saliency Map (c) Fixation Map

Figure 1.2: An example of Saliency and Fixation Map

From a theoretical perspective, saliency prediction with computational models of saliency correspond to the mechanism that results in deployment of human attention to a region in an image or video. For saliency prediction, the ground-truth data is a fixation density map (saliency map) obtained from eye movements of human observers (Fig. 1.2). Audio-visual eye tracking data is required in order to build a multi-modal saliency model, e.g. for training and evaluation, and investigating the contribution of modalities. In other words, one has to record fixations of observers while both audio and visual modalities are presented simultaneously and synchronously.

Initial efforts on the problem of video saliency predictions were limited to visual-only input. For instance, larger datasets like DHF1K [68] discard audio information during ground truth collection and ask users to look at silent videos. However, discarding audio information contrasts with our real-life behavior, where we simultaneously perceive visual and audio modalities. Psychological studies [52] , [67] indicate the impact of audio in directing the human gaze. To understand the role of audio, comprehensive eye-tracking analysis [17, 43] demonstrates that while observing a dynamic scene, the

sound will influence HVA. Audio with distinct categories, e.g., object sound, music, human voice, surrounding noise, etc., have different degrees of influence [59].

Plethora of datasets [68, 38, 14, 46, 40, 55, 35, 24, 44, 31, 62] have emerged with encouraging results and claim audio as a strong cue for multi-modal (audio-visual) video saliency prediction. These datasets have tried to capture audio-visual sequences of moving objects, dynamic natural scenes with multiple auditory conditions, short clips of actions, movie clips, unstructured user-made videos by traditional well controlled psychological experiments. Coutrot *et al.*[16], introduced a dataset claiming the improvement of only speed and accuracy of eye movement using audio-visual stimuli , compared to just audio or visual modality, according to which eyes tend to fixate more on the regions of sound sources. Benefiting from the growth of data volume, the last decade has witnessed remarkable progress in saliency detection, and plenty of methods have been proposed and achieved superior performances, especially the deep learning-based methods that have yielded a qualitative leap in performance.

A series of audio-visual saliency prediction methods followed based on the feature integration theory (FIT). i.e. individual modalities are mapped into a feature space and a saliency map is obtained by combining these features. Tavakoli *et al.* [62] proposed an audio-visual deep learning model (DAVE), where the audio and visual features are both encoded using a 3D Resnet, concatenated and sent to a decoder. Min *et al.* [45] predicts audio saliency by canonical correlation of visual and audio features and then fuses it with deep learning-based saliency models. Most state-of-the-art methods rely on 3D-CNN blocks to capture multi-level visual features. Alongside, audio features are obtained by feeding spectrograms into CNNs. [65, 74, 63] tried to model sound localization for incorporating semantically rich audio features.

Contextual analysis [62] of the audio-visual saliency model claims that the gain in performance is due to the model's ability to locate the sound source, showcasing a similar behavior to humans[22] in attending to the objects in a scene. STAViS [65] models the spatial sound source localization and obtains the feature maps that concatenate with the visual features in SUSiNet's [32] (visual saliency model) and investigates three different ways to fuse the audio modality.The audio branch is initialized by the weights trained for audio classification on the speech command database[69], whereas the video branch is initialized by the weights trained on the Kinetics Dataset[29]. Some recent efforts, have focused on face saliency, i.e. predicting the salient face in multi face videos. Liu *et al.* [35] concatenates feature from three different streams (one for faces, one for visual embedding and one for audio) and decodes saliency maps using it. In [54] they further extend their idea by bifurcating visual encoder into motion and textual features and also adding a loss function for sound localization. These works endorse audio as a significant contributing cue by reporting gains over visual-only modality that goes in line with the behavioral studies [52, 67].

Figure 1.2 shows sample frames from a multi-conversation video and the corresponding ground-truth saliency. On the first reflection of comparing visual-only and audio-visual prediction, it appears that audio plays a key role. However, when the audio-visual model (AViNet in this case) is inferred with different types of audio (zero and random), the output predictions appear identical, implying that the

audio branch is obsolete at inference. This motivated us to investigate the relevance of audio cues in conjunction with visual ones in the existing Audio-Visual Saliency Prediction models. Despite audio-visual models showcasing a gain in performance over the visual-only models, the audio information is largely ignored and AVSP models end up utilizing the visual information in a better way.

1.2 Contributions

In this thesis, we revisit these methods in audio-visual saliency and make three major observations:

1. We find that a visual-only baseline either outperforms or provides comparable performance to the state-of-the-art audio-visual saliency prediction methods.

2. We observe that the audio branch is obsolete at inference i.e., the resulting saliency maps are the same irrespective of sending zero audio, random audio, or the actual audio corresponding to the video (Fig. 1.1). We find that the observation is true for different fusion methodologies presented in the prior art.

3. Now, the interesting question is that if audio does not play any role, why does adding the audio branch lead to performance gains, at least on some datasets, as reported in previous efforts? Based on our experiments, we hypothesize that the additional branch acts as a regularizer, and the actual audio data has no role in performance improvement. We observe similar performance gains while sending randomly shuffled audio during training, which is unrelated to the video. To our surprise, a similar performance gain is observed by training an AVSP model on a visual-only dataset with random audio.

We perform comprehensive experiments to support the aforementioned claims. Our experiments comprise ten different datasets, four different fusion mechanisms, three different audio backbones, and varying experimental setups (no audio, real audio, zero audio, randomly shuffled audio, random vectors as audio). We would like to emphasize that we are not claiming any architectural novelty in this work; the goal is primarily to understand the multi-modal learning and provide essential cues that will help better design and evaluation of future audio-visual saliency prediction models. This work questions the role of audio in current end-to-end trained deep learning saliency prediction methods. It motivates reconsideration of the complex architectures for audio-visual saliency prediction by demonstrating that the simpler alternatives work equally well. It encourages a more rigorous evaluation of the saliency prediction methods in the multi-modal setting. And finally, it highlights the limitations of the current efforts and motivates exploration of ways to actually exploit the audio information for the task of saliency prediction.

1.3 Thesis Outline/Organization

The rest of the thesis is organized as follows:

- Chapter 2 discusses the related work(psychological studies and computation modeling) and contemporary architectures of multi-modal Saliency Prediction that demonstrate the performance gains by adding audio.
- Chapter 3 discusses our methodology to analyze the role of audio in existing state-of-the-art saliency prediction models and validate the efficacy of audio branch. We hypothesize that the audio module acts as a regularizer and produce experimental validation for the same.
- Chapter 4 presents the datasets, training procedure and evaluation metrics used.
- Chapter 5 focuses on extensive experiments performed to examine the role of audio in current AVSP models. The study is carried out on ten different audio-visual saliency datasets and also attempts to investigate the cause for incremental gains in current AVSP models over the visual-only models.
- Chapter 6 summarizes our work and presents the concluding thoughts with future direction by highlighting the limitations of the current efforts and exploring ways to exploit the audio information for the task of saliency prediction.

Chapter 2

Related Works

Humans are intelligent multi-sensory creatures, capable of spotting and focusing on certain parts of audio or visual stimuli in a cluttered environment (i.e., have attentional behavior). It is, hence, unsurprising that inspired by such observations psychologists and neuroscientists have been studying attentional mechanisms underlying auditory, visual, and auditory-visual attention. Several decades of research on attention mechanisms have amassed a rich literature on the topic. The last decade has witnessed remarkable progress in saliency detection, and plenty of methods have been proposed and achieved superior performances, especially the deep learning-based methods that have yielded a qualitative leap in performance. Covering the whole literature is, thus, beyond the scope of this thesis. Instead, to stress the influence of audio on multi-modal attention models, the following sections provide a brief account of relevant studies.

2.1 The role of audio in HVA

Hearing and sight, which mainly are relied on by humans when perceiving the world, occupy a considerable portion of the received external information[37]. Our brain comprehensively understands the environments by integrating these multi-modal signals with different forms and physical characteristics. For example, in the cocktail party scenario with many speakers, we can locate the one of interest and enhance the received speech with the aid of his/her lip motion. Hence, audio-visual learning is essential to our pursuit of human-like machine perception ability. Its purpose is to explore computational approaches that learn from audio-visual data.

The span of behavioral studies on audio-visual attention is broad and covers a wide range of experiments on primates and humans. In such experiments, an observer is often presented with a stimulus and his neural and/or behavioral responses are recorded (e.g.using single unit recording, brain imaging, or eye tracking). Audio matters in Human Visual Attention (HVA). Papai et al. [51] points everyday examples like stopping in our tracks at the sudden car honk while absentmindedly crossing a street. However, is it still crucial while watching a video with a monaural audio? Numerous studies suggest that it is. Chen *et al.* [11] captured eye gaze on images with no audio, coherent audio and incoherent

audio. They found that coherent audio information is an important cue for enhancing the feature-specific response to the target object. Eye tracking experiments in previous works [14, 64] also verify the impact of audio signal on human attention. Audio-Visual interactions like The McGurk effect[41] showed how mismatched auditory and visual stimuli when combined give a changed perception. Burg *et al.* [66] also showed that non-spatial temporal audio signal when interacted with a visual event made the visual target more salient. The work in [15, 39] finds noticeable differences in spatial distributions of visual attention on same video content, when viewed with and without audio. Eye tracking experiments by Coutrot *et al.* [15] further suggest that in conversational video, increasing saliency of speakers' face greatly improves the model prediction. Other similar studies [17, 18, 58] also confirm the impact of soundtrack on gaze while watching videos. Our work investigates if a similar behaviour is observed in deep learning based saliency prediction models.

2.2 Computational Saliency Prediction

Figure 2.1: Frames (top row), Ground Truth Saliency (bottom row). Same object moving in a scene in a video clip usually attracts visual attention.

Videos contain spatial information in frames and temporal information between frames. In videos, human attention is guided by low-level cues and semantic context in a single frame, as well as by relations of features in frames. For example, the same object moving in a scene in a video clip usually attracts visual attention(Fig. 2.1). Consequently, it is crucial for video saliency prediction (VSP) to synchronously exploit spatial and temporal information.

2.2.1 Generic Image and Video Saliency Prediction

Early methods related to VSP mainly explored static and motion information using low-level hand-crafted features such as intensity, color, orientation channels, etc. However, these were not powerful enough to model dynamic saliency, after which several deep learning-based VSP models emerged. Initial deep learning-based saliency prediction methods were limited to visual information. Bak *et al.* [2] proposed a two-stream network, using convolutional backbones, with RGB images and optical flow

maps as inputs, respectively, to extract spatiotemporal information and fused them for final saliency inference. Recent methods can be largely classified into two categories :

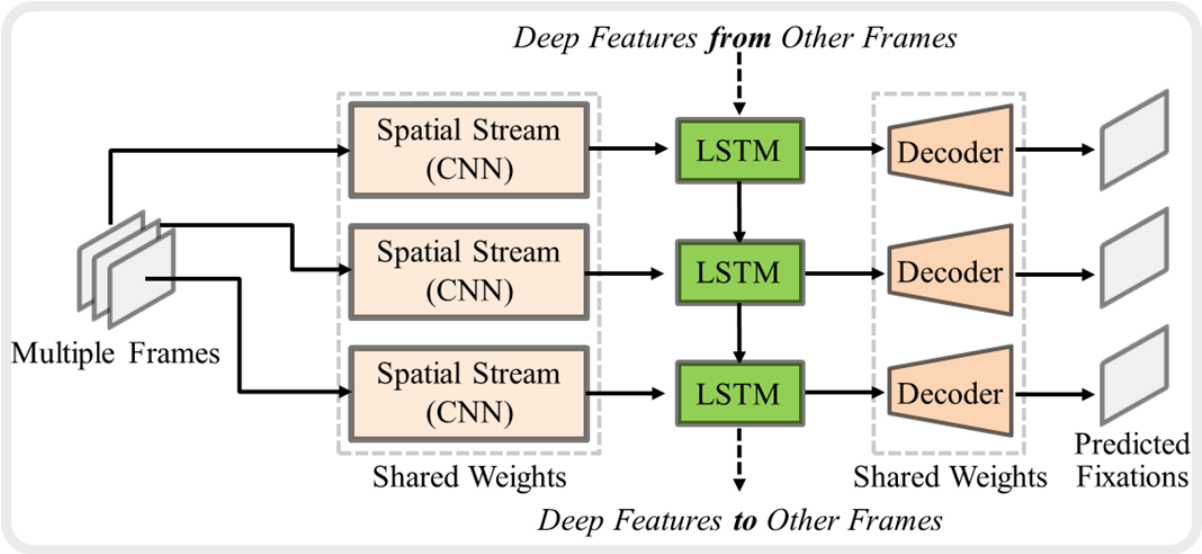


Figure 2.2: Method pipeline of the long short-term memory (LSTM) based approaches which usually follow the single stream methodology.

- *The LSTM based models*, which build on image-based saliency and aggregate frame-wise prediction using an LSTM [19, 70] (Fig. 2.2). LSTMs were adopted to extend the temporal information as opposed to the optical flow network, which only considers the temporal relationships between subsequent frames. This approach usually uses convolutional neural networks (CNN) to compute deep spatial features for each frame. Then, to sense temporal information, all deep features computed individually via CNN are fed into the input gate of LSTM. Finally, a decoder is applied to produce the pixel-wise fixation prediction.

  Gorji *et al.* [23] employed multi-stream LSTMs and merged each static saliency map for VSP. DeepVS [28] leveraged sub-networks of object and motion to extract intra-frame saliency information, and modeled temporal correlation between frames by convLSTMs. Besides, ACLNet [68] adopted an attention module in a CNN-LSTM structure, which was supervised with image SP datasets. STRA-Net [33] proposed a two-stream model, by employing dense residual cross-connection to enrich interactions between motion and appearance stream during feature extraction, and incorporated an attention mechanism to enhance the spatio-temporal information. SalEMA [34] modified the static SP model by using an exponential moving average instead of LSTM for feature fusion in the temporal domain, resulting in a low-parametric architecture. Later, Zhang *et al.* [73] utilized spatial and channel attention to select and re-weight spatiotemporal information, and employed an attentive convLSTM to model relations between frames. SalSAC

[70] designed a correlation-based convLSTM for VSP, in which adjacent frames were weighted according to the similarity between them i.e. balancing the saliency alteration by the change of image characteristics in consecutive frames. The main drawback with all LSTM-based approaches is that they overlay temporal information on top of spatial information rather than utilizing both kinds of information simultaneously, which is essential for VSD.

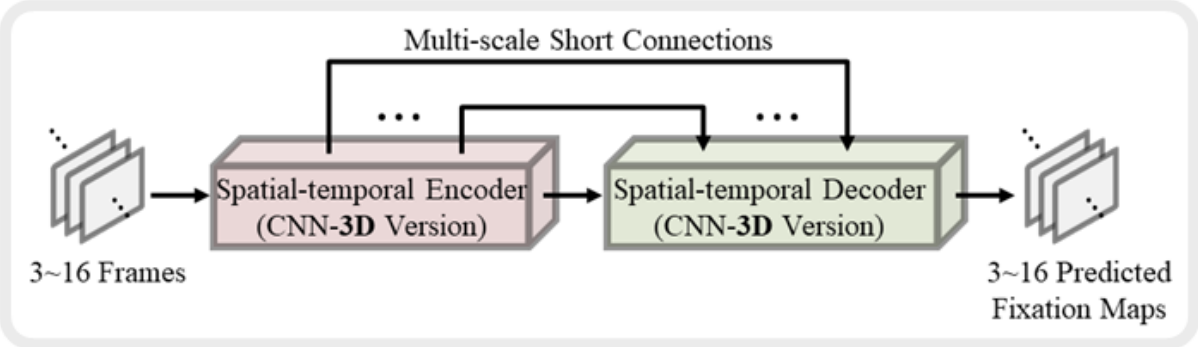


Figure 2.3: Method pipeline of the 3D convolution-based approaches and the major highlight of these approaches is their capability of sensing both spatial and temporal information

- *The 3Dconv based models*, which rely on action detection networks as their backbone and primarily use 3D convolutional layers in the encoder and the decoder [42, 27](Fig. 2.3). Compared with LSTM-based architectures, which process spatial and temporal information separately, a 3D network encodes and decodes spatio-temporal information in a collective way.

  TASEDNet [42] proposed a 3D fully-convolutional encoder-decoder network for VSP by adopting S3D network as encoder and the decoder uses 3D deconvolution and unpooling so as to continuously enlarge the image to obtain the saliency map, and achieved promising performance by considering the influence of space, time and scale. Moreover, ViNet [27] designed a UNet-like encoder-decoder network based on a 3D backbone, in which features from multiple levels were upsampled with trilinear upsampling and concatenated along the temporal channel. In HD^2S [3], multilevel features from a 3D encoder are separately decoded to obtain single-channel conspicuity maps, and integrates all the decoded feature maps to obtain the final saliency map. Besides, TSFP-Net [8] employs a feature pyramid structure with top-down feature integration on a 3D convolutional backbone, and combines the multi-level spatiotemporal features to reason the saliency result for a video frame.

Needless to say, the 3Dconv-based architectures borrow common ideas from deep learning research like using features from different hierarchies, skip connections, transfer learning, multi-branch architectures, UNet like encoder-decoder [56] etc. Most methods, first train the saliency prediction model using DHF1K dataset and then fine-tune it on other datasets. The current state-of-the-art landscape is domi-

nated by 3DConv-based architectures. We rely primarily on the ViNet [27] model for our experiments, owing to its simplicity and decent performance.

2.2.2 Audio-Visual Saliency Prediction

Some recent architectures have begun to explore the impact of multi-modal information on saliency(mainly Audio-Visual). The Audio-Visual saliency prediction methods fuse the visual branch with audio information. Several fusion methodologies have been studied in prior art. Tavakoli *et al.* [62] uses a 3D Resnet to encode both visual and audio information. They employ a simple concatenation operation on the encoded features. Chen *et al.* [9] also use concatenation operations, while using features from different visual hierarchies. STAViS [65] fuses the audio features by performing a spatial sound source localization onto the SUSiNet [32] visual encoder. They employ three different fusion methodologies namely cosine similarity, weighted inner product and bilinear transformation. Zhu *et al.* [74] employ a linear weighted fusion of audio and visual saliency maps. The audio saliency maps are computed using canonical correlation of visual and audio features. Jain *et al.* [27] experiment with similar fusion methods to [32] on the ViNet backbone.

Some saliency prediction efforts have focused on conversational multi-face videos. Liu *et al.* [35, 54] employ multi-stream end-to-end trainable deep learning architectures. They propose a large scale MVVA dataset allowing efficient training. Several non deep learning methods have also been explored [15].

Most of the aforementioned audio-visual saliency prediction methods claim that fusing audio leads to noticeable performance gains. Jain *et al.* [27] were the first to question this claim, by showing that an optimally trained visual backbone, can match the performance of audio-visual methods. They demonstrate that the performance gains by adding audio are not statistically significant. We make a more comprehensive effort in this direction, performing experiments with different audio backbones and a variety of fusion methodologies. Our work also provides insights on why performance gains are observed by fusing audio in training, their role at inference and a comparison to other regularization techniques.

Chapter 3

Methodology

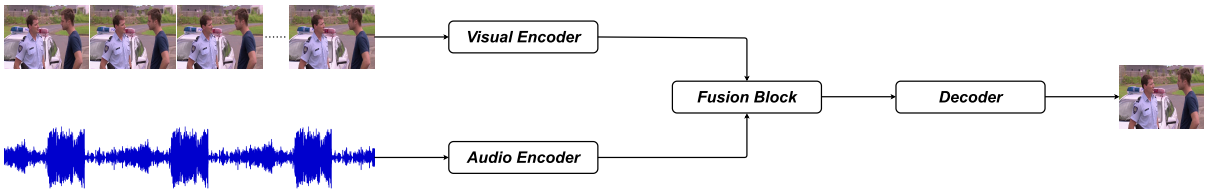


Figure 3.1: Audio-Visual Saliency Prediction model in general.

We analyze role of audio in existing state-of-the-art saliency prediction models (Section 3.1) and validate the efficacy of audio branch. We then evaluate the effectiveness of different encoding (Section 3.2) and fusing strategies (Section 3.3) towards the same. Furthermore, we corroborate the underlying cause for incremental gains in all existing AVSP models. We hypothesize that the audio module acts as a regularizer (Section 3.4) and produce experimental validation for the same.

3.1 Audio-Visual Saliency models

Existing deep audio-visual saliency models can be interpreted as an encoder-decoder framework (Fig. 3.1). For this study, we choose STAViS[65] and AViNet[27] networks that fuse spatio-temporal visual and auditory information to obtain a final saliency map.

3.1.1 STAViS

STAViS is an end-to-end spatio-temporal audiovisual saliency network that combines visual saliency and auditory features, and learns to appropriately localize sound source by fusing the two saliencies to obtain a final saliency map. We train STAViS [65] that extends the SusiNet [32] visual saliency model by fusing an audio modality (Fig. 3.2). The visual branch consists of spatio-temporal module based on 3D-ResNet Blocks pre-trained on Kinetics-400 dataset [7]. This is followed by a Deeply Supervised Attention Module (DSAM) where hierarchical features from multiple layers of visual encoder are passed

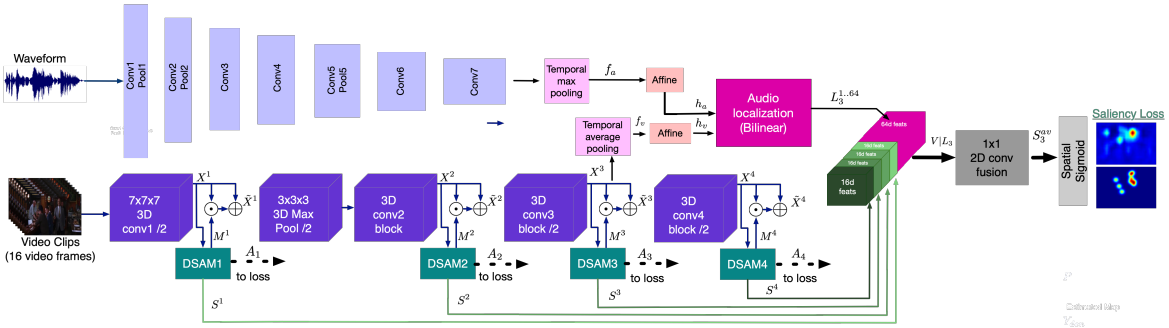


Figure 3.2: STAViS architecture: the Spatio-Temporal Audio-visual network is based on the ResNet architecture and has a spatio-temporal visual branch, auditory branch and their fusion.

and concatenated at highest level, i.e., element-wise multiplication of the output of each block(feature map) and the attention map, to enhance the most salient regions of the feature maps. In parallel, semantically rich audio features are obtained using SoundNet [1] , a state-of-the-art CNN for acoustic event classification, and then combined with visual encoder feature map to obtain a final saliency map. The pre-processing is done similar to [65]. The model is trained on a weighted combination of three loss functions :

$$\mathcal{L}^i_{sal}(W) = w_1\mathcal{L}^i_{CE} + w_2\mathcal{L}^i_{CC} + w_3\mathcal{L}^i_{NSS}$$

where $\mathcal{L}^i_{CE}, \mathcal{L}^i_{CC}$ and $\mathcal{L}^i_{NSS}$ are Cross-Entropy, Linear Correlation Coefficient and Normalized Scanpath Saliency loss function and w_1, w_2, w_3 are the weights of each loss type respectively. For our case, w_1, w_2, w_3 empirically decided as $0.1, 2, 1$ respectively by [36] is used.

3.1.2 AViNet

We train AViNet (Fig. 3.3), a U-Net like encoder-decoder network with a visual branch based on a S3D [71] backbone pre-trained on Kinetics-400 action recognition dataset [7] and lacks explicit inputs such as optical flow, or additional modules for detecting motion, object, attention, etc. Features from multiple levels are upsampled with trilinear interpolation and combined along the temporal channel. Inspired by STAViS, the SoundNet[1] module is used as an auditory feature extractor. The audio features are fused with visual features by simple concatenation and Bilinear techniques. Inputs are processed similar to [27]. The model is trained on KLdiv loss:

$$KLdiv(A, B) = \sum_i B_i \log(\epsilon + \frac{B_i}{A_i + \epsilon})$$

here A,B are predicted and ground truth maps respectively and ϵ is a regularization term.

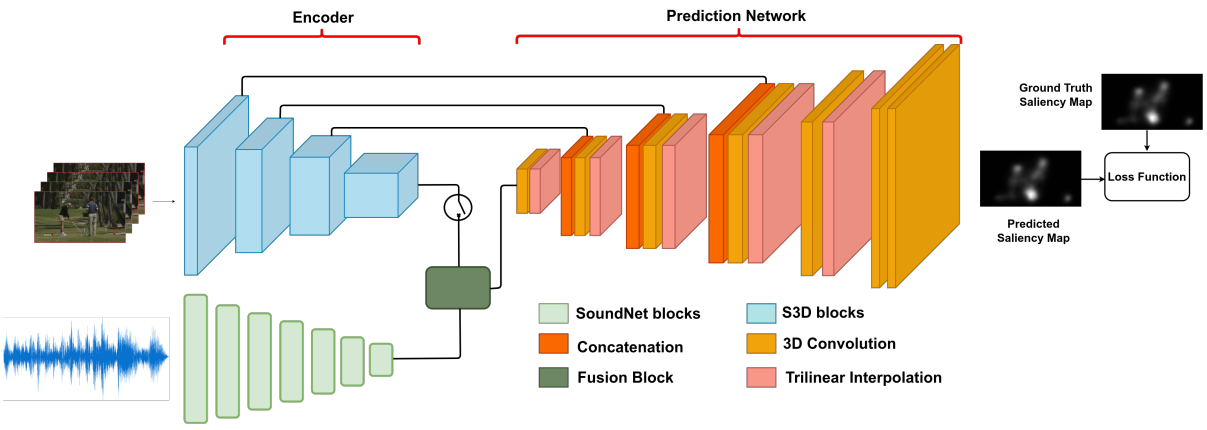


Figure 3.3: Overview of the AViNet architecture. ViNet is the architecture that results from removing the audio branch.

3.2 Audio Modules

To analyze the role of audio, we perform ablation across three different audio modules, *i.e.*SoundNet[1], VGG-Vox[12] and AVID[48]. These modules have shown significant performance in diverse correlated audio-visual tasks.

3.2.1 SoundNet

For sound representation, we employ SoundNet[1] (Fig. 3.4) to leverage visual and sound synchronized information in the videos. It uses a student-teacher model that transfers discriminative visual information from well-established visual recognition models, employing a massive source of unlabelled video as a bridge. High-level feature embeddings are then extracted from the seventh layer of SoundNet with dimension of 1024×3, followed by temporal max-pooling layer. This module is fine-tuned by end-to-end training for our AVSP task.

3.2.2 VGG-Vox

We also employ VGG-Vox[12] (Table 3.1) as an audio module, which is a modified version of a speaker recognition network VGG-M. The input to this network is a short-term amplitude spectrogram extracted from raw audio (with same duration as of input video) using a 512-point FFT , resulting in a spectrogram of size 512×300. Each frequency bin of the spectrogram is normalized and fed to the audio module, which aggregates frame-level feature vectors to obtain a fixed-length utterance-level embedding of dimension 4096. The VGG-Vox model pretrained on Voxceleb2[12] dataset is fine-tuned for our task by end-to-end training of AVSP model.

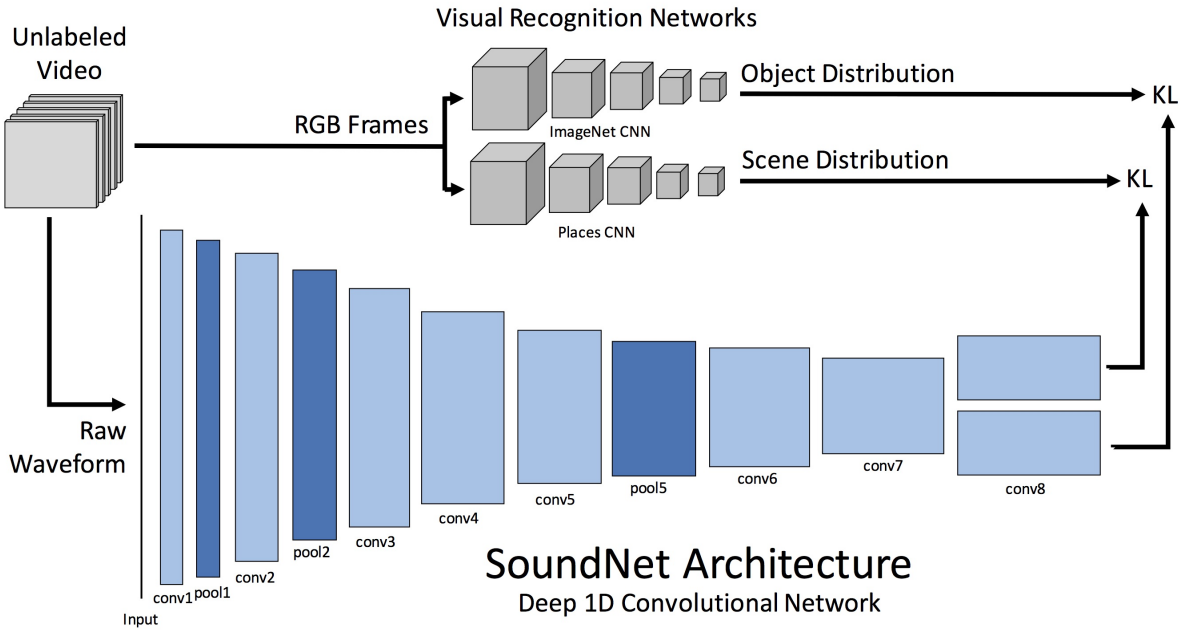


Figure 3.4: SoundNet Architecture overview

Layer	Support	Filter dim.	# filts.	Stride	Datasize
conv1	7×7	1	96	2×2	254×148
mpool1	3×3	-	-	2×2	126×73
conv2	5×5	96	256	2×2	62×36
mpool2	3×3	-	256	2×2	30×17
conv3	3×3	256	384	1×1	30×17
conv4	3×3	384	256	1×1	30×17
conv5	3×3	256	256	1×1	30×17
mpool5	5×3	-	-	3×2	9×8
fc6	9×1	256	4096	1×1	1×8
apool6	$1 \times n$	-	-	1×1	1×1
fc7	1×1	4096	1024	1×1	1×1
fc8	1×1	1024	1251	1×1	1×1

Table 3.1: VGG-Vox architecture overview.

3.2.3 AVID

Furthermore to verify the role of audio in the aforementioned task, we employ AVID[48] module to learn audio representations by using contrastive learning for cross-modal discrimination between the two modalities in a self-supervised manner. Audio is processed by sampling with a time-frame of input video sequence, and a log spectrogram of size 100×129 is obtained where 100 is the number of time

steps, and 129 is the number of frequency bands choosen in our case. This spectrogram is then passed through 9 layers of 2D ConvNet and projected to 128 dimensions using a multi-layer perceptron (MLP) composed of 3 fully connected layers with 512 hidden units. We fine-tune the pretrained AVID model by end-to-end training of resulting AVSP model.

3.3 Fusion of Multi-Modalities

We exploit different fusion techniques (Fig. 3.5) for our analysis, owing to their ability to generalize well across multiple domains, thereby leveraging multi-modal information.

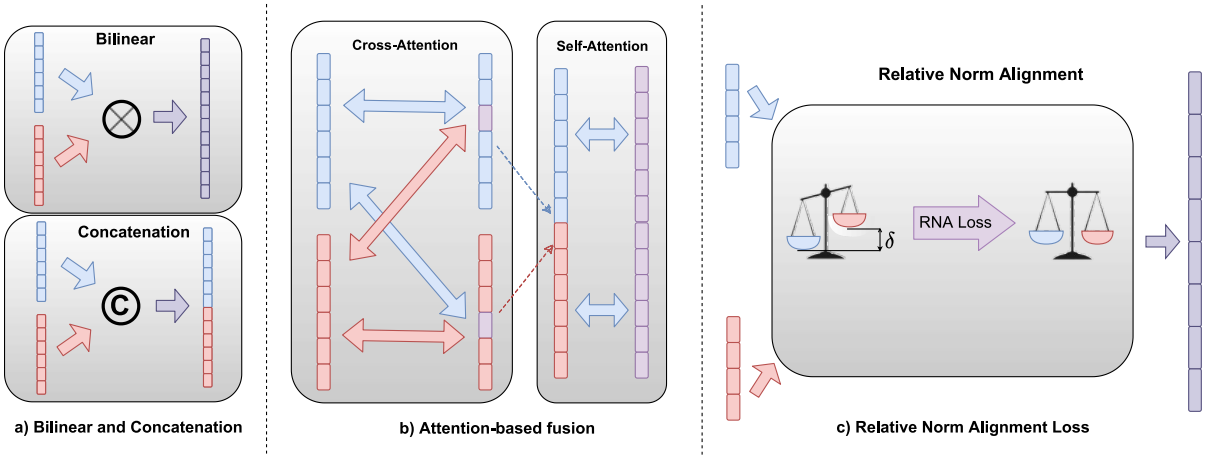


Figure 3.5: Different Fusion Techniques

3.3.1 Bi-linear Fusion and Concatenation

Inspired by the recent works of audio-visual fusion for Saliency, we apply bi-linear fusion and concatenation techniques used in [65, 27]. Bilinear fusion method captures the pairwise interactions across the feature dimensions. For bi-linear fusion, after applying the max-pool layer in the last layer of encoder, the visual and audio features are flattened to form a vector of dimension $x_v \in \mathbb{R}^{1024 \times x_o}$ and $x_a \in \mathbb{R}^{1024 \times y_o}$ respectively. It is formulated as :

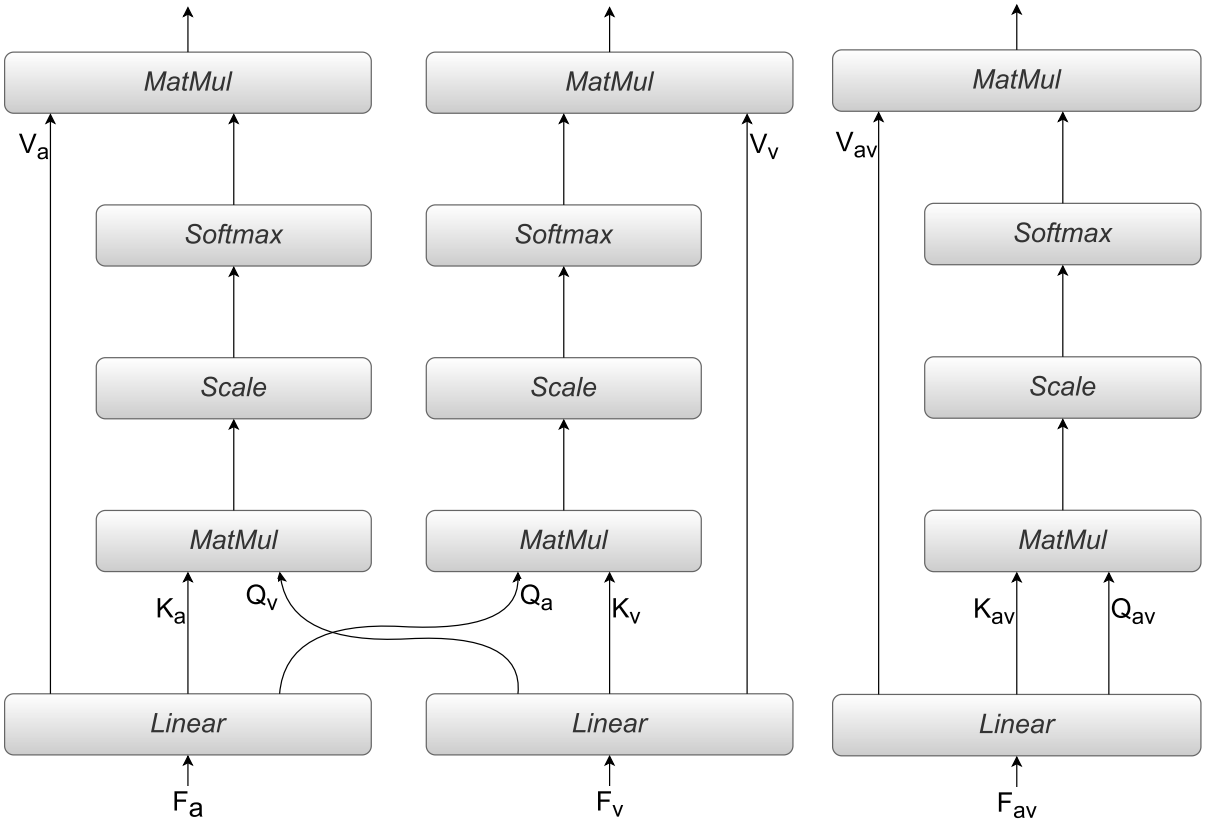
$$y = x_v^T A x_a + b$$

where $A \in \mathbb{R}^{x_o \times x \times y_o}$ and $b \in \mathbb{R}^{x \times 1}$ are parameters and x is the desired output dimension.

We also performed our experiments with a simple concatenation technique used in [65, 27]. To match the dimensions for concatenation , audio features are repeated and combined based on the number of channels. This fusion is followed by a 1×1 convolution to reduce the channel dimension.

3.3.2 Self-Attention and Cross Attention

Instantaneous sound content and activities in the scene may not always be precisely time-aligned, thereby causing the two modalities to possess distinct dynamics. Motivated by [61], proposed initially for speaker detection, we employ a cross-attention and self-attention module to capture the dynamic visual-audio interaction along the temporal dimension (Fig. 3.6). Attention based mechanism synergistically combines the two modalities. Cross attention ensures that attention features from one modality are used to highlight the features of other modality , thereby capturing the inter-modality interaction. Subsequently, self-attention is applied to capture long-term temporal dependencies, where the attention mask highlights its own spectral features. Self-attention mechanisms can extract features globally, which is more potent for modeling the long-range correlations between video frames in temporal sequences. The cross-attention module projects the processed audio and visual features on the same feature space.



(a) cross-attention network

(b) self-attention network

Figure 3.6: Fig. 3.6a and Fig. 3.6b represents the attention layer in Cross and Self-Attention network respectively. Considering the audio embeddings F_a as the source and the visual features F_v as the target, we generate audio attention feature $F_{a \to v}$ as the output. In a similar way, visual attention feature $F_{v \to a}$ is generated

As audio and visual flow each has its own dynamics, they need to be more precisely time aligned. So, we employ cross-attention networks along the temporal dimension to dynamically describe such audio-visual interaction. The core part of the cross-attention network is the attention layer, which is shown in Fig. 3.6a. The inputs are the vectors of query (Q_a, Q_v),key (K_a, K_v), and value (V_a, V_v) from audio and visual embeddings respectively, projected by a linear layer. The outputs are the audio attention feature $F_{a\to v}$ and visual attention feature $F_{v\to a}$ as formulated in Eq. (3.1) and (3.2), where d denotes the dimension of Q, K and V :

$$F_{a\to v} = softmax(\frac{Q_v K_a^T}{\sqrt{d}})V_a \tag{3.1}$$

$$F_{v\to a} = softmax(\frac{Q_a K_v^T}{\sqrt{d}})V_v \tag{3.2}$$

As formulated in Eq. (3.1) and (3.2), to learn the interacted new audio feature $F_{a\to v}$, the attention layer applies F_v as the target sequence to generate query, and F_a as the source sequence to generate key and value, and to learn $F_{v\to a}$ vice-versa. The feed-forward layer follows the attention layer. Residual connection and layer normalization are also applied after these two layers to generate the whole cross-modal attention network. The outputs are concatenated together along the temporal direction.

A self-attention network is applied after the cross-attention network to model the audio-visual utterance-level temporal information. As shown in Fig. 3.6b , this network is similar to the cross-attention network except that now the query Q_{av}, key K_{av}, and value V_{av} in the attention layer all come from the joint audio-visual feature F_{av}

3.3.3 RNA Loss

Though multiple modalities may provide additional information, CNNs' ability to effectively extract valuable information from them is limited due to one modality being *"privileged"* over the other during training, limiting its generalization ability. To this end, Planamente *et al.* [53] brought into light the problem of *"norm unbalance"* and reported L2-norm as the metric to measure the unbalance between the information content of the training modalities. The mean-feature-norm distance (δ) between the two modality norms f^v and f^a can be computed as:

$$\delta(h(x_i^v), h(x_i^a)) = |\mathbb{E}[h(X^v)] - \mathbb{E}[h(X^a)]|$$

where $\mathbb{E}[h(X^m)]$ corresponds to the mean features norm for each modality. Figure 3.7 illustrates the norm $h(x_i^v)$ of the i^{th} visual sample and $h(x_i^a)$ of the i^{th} audio sample, by means of segments of different lengths arranged in a radial pattern.The mean feature norm of the k^{th} modality is represented by the radius of the two circumferences, and δ is represented as their difference. The objective is to minimize the δ distance by means of a new loss function, which aims to align the mean feature norms of the two modalities. In the case of Audio-Visual modality, the objective is to minimize the difference between the radius of respective norms forcing them to lie on a hyper-sphere of a fixed radius. Planamente *et*

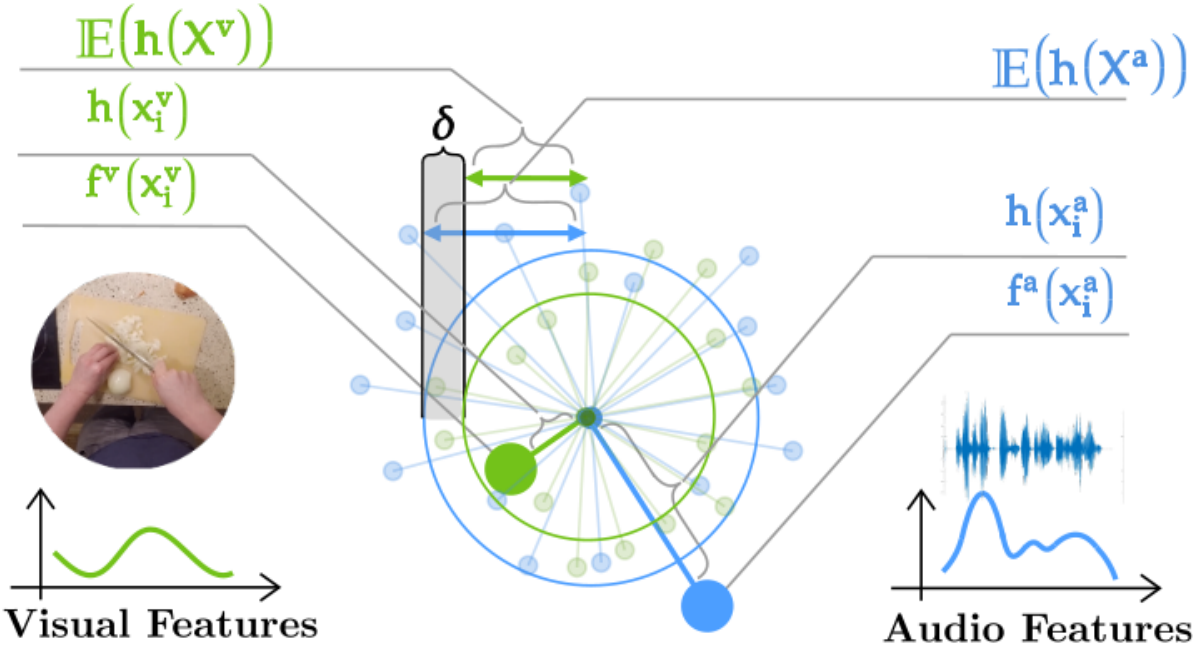


Figure 3.7: The norm $h(x_i^v)$ of the i^{th} visual sample (left) and $h(x_i^a)$ of the i^{th} audio sample (right) are represented, by means of segments of different lengths. The radius of the two circumferences represents the mean feature norm of the two modalities, and δ their discrepancy. By minimizing δ, audio and visual feature norms are induced to be the same.

al. [53] proposed a Relative Norm Alignment (RNA) loss that aims to align the mean feature norms of the two modalities from various source data, resulting in domain-invariant audio-visual features. RNA loss can be defined as :

$$\mathcal{L}_{RNA} = \left(\frac{\mathbb{E}[h(X^v)]}{\mathbb{E}[h(X^a)]} - 1 \right)^2,$$

where $\mathbb{E}[h(X^m)] = \frac{1}{N} \sum_{x_i^m \in \chi^m} h(x_i^m)$, for the m^{th} modality and N denotes the number of samples of the set $\chi^m = \{x_1^m, \ldots, x_N^m\}$. In order to induce an optimal equilibrium between the two embeddings, the dividend/divisor structure is adjusted to encourage a relative balance between the norm of the two modalities. Furthermore, the square of the difference pushes the network to take larger steps when the ratio of the two modality norms is too high resulting in faster convergence.

3.4 Regularization over Visual-Only Models

Dropout [60] is a regularization technique to ameliorate over-fitting in neural networks. Specifically, during the training phase, dropout randomly discard nodes with a given probability. In this way, the network can be hypothesized as an ensemble of small sub-networks, thus achieving a good regularization

effect. For our visual only model, we use high dropout of 85%. (value of Dropout is decided empirically based on Table 5.6)

Chapter 4

Experiments

4.1 Dataset

We carry out the tests and comparisons on :

- *three most popular visual-only saliency datasets* - DHF1K, Hollywood-2, and UCF Sports
- *six audio-visual saliency datasets* - AVAD, Coutrot1, Coutrot2, DIEM, ETMD, SumMe
- *two multi-face datasets* - Coutrot2, MVVA.

4.1.1 Visual-only Datasets

During the ground-truth collection of visual-only datasets, the audio was discarded. Therefore, users were asked to look at *silent videos*.

4.1.1.1 DHF1K

DHF1K [68] is a large dataset with diverse content and variable length comprising 1000 videos split into 600, 100, and 300 as training, validation, and testing sets. Each video is 30 fps with 640x360 spatial resolution, and eye-tracking data annotated by 17 observers. The dataset is mainly classified into 7 categories: humans (daily activities, sports, social activities, and art), animals, artifacts, and scenery. The ground truths of testing videos are held out for evaluation on the benchmark website .

4.1.1.2 Hollywood-2

Hollywood-2 [40] is the largest dataset in terms of the number of videos, consisting of 1707 action videos from the Hollywood-2 action recognition dataset with eye-tracking data annotated by 19 observers. The dataset has short video sequences from a set of 69 Hollywood movies containing 12 different human action classes, ranging from answering the phone, eating, driving, running, etc. We use the standard split of 823 training videos and 884 test videos.

4.1.1.3 UCF Sports

UCF Sports [55] dataset consists of a set of actions collected from various sports which are typically featured on broadcast television channels. The dataset includes a total of 150 sequences with a resolution of 720 x 480. It includes 10 actions, i.e., diving, golf swing, kicking, lifting, riding a horse, running, skateboarding, swing-bench, swing-side, and walking. We use a standard split with 103 videos for training, and 47 videos for testing.

4.1.2 Audio-Visual Datasets

4.1.2.1 DIEM

DIEM [46] consists of 84 videos with varying genres based on g advertisements, documentaries, game trailers, movie trailers, music videos, news clips, and time-lapse footage. The eye-tracking data are annotated by about 50 observers in a free-viewing manner. We use a standard split with 20 videos for testing and the remaining videos for training.

4.1.2.2 Coutrot

Coutrot [14, 38] databases are divided into Coutrot1 and Coutrot2.

- Coutrot1 contains 60 clips with dynamic natural scenes of four visual categories: one/several moving objects, landscapes, and faces.

- Coutrot2 contains 15 clips of 4 persons in a meeting. Videos have a resolution of 720 x 576 pixels and a frame rate of 25. The corresponding eye-tracking data are annotated by 70 observers.

4.1.2.3 SumMe

SumMe [24] dataset contains 25 unstructured videos, i.e., mostly user-made videos and their corresponding multiple-human created summaries, which were acquired in a controlled psychological experiment. The corresponding eye-tracking data are annotated by 10 observers.

4.1.2.4 AVAD

AVAD [44] dataset comprises 45 short clips of 5-10 sec duration with several action scenes, like dancing, guitar playing, birds singing, etc. The corresponding eye-tracking data are annotated by 16 observers.

4.1.2.5 ETMD

ETMD [31] dataset consists of 12 videos from 6 different Hollywood movies. The corresponding eye-tracking data are annotated by 10 observers.

4.1.3 Multi-Face Datasets

4.1.3.1 MVVA

MVVA [35] dataset consists of 300 dubbed Multiple-face Videos. The corresponding eye-tracking data are annotated by 34 observers. During the eye-tracking experiment, both video and audio were presented to the annotators. A random split of 240 videos for training and 60 videos for testing is used.

4.1.3.2 Coutrot2

Discussed in Section 4.1.2.2

4.2 Training procedure

For training AViNet, a similar training procedure is incorporated, as discussed in [27]. 32 consecutive frames are randomly selected from each clip of the dataset with their corresponding audio stream. Each frame is resized to 224×384 and trained with a batch size of 8. The optimizer used is Adam with an initial learning rate of 1e-4 and Kullback-Leibler divergence as the loss function. The network is initially trained on DHF1K dataset with corresponding validation data used for early stopping. The network with pre-trained weights of DHF1K dataset is fine-tuned for all other datasets with their respective validation datasets being used for early stopping.
For a fair comparison, the training procedure of STAViS is adopted as discussed in [65]. The network takes 16 consecutive frames as input with a resolution of 112 X 112 and is trained with a batch size of 128 with their corresponding audio stream. A random flipping data augmentation technique is applied during training. The optimizer used is SGD with a momentum of 0.9, dampening factor of 0.9, weight decay of 1e-5, and learning rate set to 1e-2. The loss function is a weighted combination of cross-entropy loss, linear correlation coefficient (CC), and normalized scanpath saliency(NSS).

4.3 Evaluation Metrics

We evaluated our task on distribution-based and location-based metrics [6] (Table 4.1). Distribution-based metrics compute the similarity between predicted and ground truth distributions (assuming that the ground truth fixation locations are sampled from an underlying probability distribution). We chose KLDiv, Similarity, and Correlation(CC) for distribution-based analysis. The location-based metrics

measure the accuracy of saliency models at predicting discrete fixation locations. NSS and AUC metrics are chosen as location-based metrics in our analysis.

Table 4.1: Distribution-based metrics consider both predicted saliency maps and ground truth fixation maps as continuous distributions, while Location-based metrics account for saliency map values at discrete fixation locations. High values for similarity metrics and low values for dissimilarity metrics are characteristics of good saliency models.

Metrics	**Location-based**	**Distribution-based**
Similarity	AUC, sAUC , NSS	SIM, CC
Dissimilarity	X	KLDiv

4.3.1 Distribution-based metrics

4.3.1.1 KLDiv

Kullback-Leibler divergence treats saliency maps as probability distributions and measures the loss of information between the predicted saliency map and the ground truth saliency map by the difference between two distributions

$$KLdiv(A, B) = \sum_i B_i \log(\epsilon + \frac{B_i}{A_i + \epsilon})$$

here A,B are predicted and ground truth maps respectively and ϵ is a regularization term. Lower value indicate better approximation of predicted saliency map with the ground truth saliency map.

4.3.1.2 CC

The Pearson's Correlation Coefficient(CC) measures the correlation between two variables. CC can be used to interpret saliency and fixation maps, A and B^D as random variables and measure the linear relationship between them.

$$CC(A, B^D) = \frac{\sigma(A, B^D)}{\sigma(A) \times \sigma(B^D)}$$

where $\sigma(A, B^D)$ represents covariance between A and B^D. It is a symmetric metric and penalizes false positives and negatives equally. It is invariant to linear transformations. The pixels where both predicted and ground truth saliency maps have similar values gives high CC values. CC value can vary from -1 to 1.

4.3.1.3 SIM

The Similarity metric (SIM) measures the similarity between two distributions, viewed as histograms (so, also referred to as histogram intersection). SIM is computed as the sum of the minimum values at each pixel, after normalizing the input maps. Given a saliency map A and a continuous fixation map B^D :

$$SIM(A, B^D) = \sum_i \min(A_i, B_i^D)$$

where $\sum_i A_i = \sum_i B_i^D = 1$, by iterating over discrete pixel locations i. A similarity score of one indicates that the maps are identical, and zero indicates that they are entirely dissimilar.

4.3.2 Location-based metrics

4.3.2.1 NSS

Normalized Scanpath Saliency (NSS) aims to quantify the saliency map values at the fixated locations and normalize them with the predicted map variance. Given a saliency map P and a binary map of fixation locations Q^B:

$$NSS(P, Q^B) = \frac{1}{N} \sum_i \bar{P}_i \times Q_i^B$$

$$\text{where } N = \sum_i Q_i^B \text{ and } \bar{P} = \frac{P - \mu(P)}{\sigma(P)}$$

Here, i indexes the i^{th} pixel, $\bar{P}$ is the normalized saliency map, and N is the total number of fixated pixels. NSS is sensitive to false positives, general monotonic transformations and relative differences in saliency across the image. However,since the mean saliency value is subtracted during computation, NSS is invariant to linear transformations. High-valued prediction at fixated locations gives a higher NSS score.

4.3.2.2 AUC-Judd

The Area Under the ROC Curve (AUC) is the most widely used metric for evaluating saliency maps. The saliency map is treated as a binary classifier of fixations at various threshold values (level sets), and a ROC curve is swept out by measuring the true and false positive rates under each binary classifier. Its value can vary from 0 to 1.

4.3.2.3 sAUC

It penalizes models that include the bias of the emergence of a central Gaussian distribution when averaging fixations over many images by sampling negatives from fixation locations from other images

instead of uniformly at random. Its value can vary from zero to one and higher the value better are the predictions.

Chapter 5

Results and Discussions

We conducted a comprehensive series of experiments to analyze the role of audio in audio-visual saliency prediction (AVSP) models. Our investigation involved evaluating the performance of AVSP models on ten different audio-visual saliency datasets. Additionally, we explored the underlying reasons for the observed incremental gains of AVSP models over visual-only models, aiming to gain insights into the contributing factors for the improved performance of AVSP models.

5.1 Audio-visual Dataset

5.1.1 Role of Audio in SOTA models

To analyze the influence of audio in AVSP models, we conduct a simple experiment by setting the sound signal to zero (a silent sound), and random (a random noise) at inference. From Table 5.1, we find that the model inferred with different sound signals gives notably similar performance, thus showing an agnostic behavior of both the SOTA models with audio on all the audio-visual datasets.

For instance, when examining the Coutrout 2 dataset, we observed that AViNet consistently demonstrated the highest gains in performance compared to ViNet. Intriguingly, even when the audio input was set to zero or a random audio vector, AViNet's performance remained unchanged. These results suggest that AViNet may not effectively utilize audio information for saliency prediction. Despite the absence of meaningful audio input, AViNet's performance was on par with its performance when provided with actual audio input, indicating a lack of sensitivity to audio information.

This behaviour suggest that the SOTA models are unable to utilize audio module at it's best and limits the performance of AVSP models. Motivated by these findings, we aimed to explore different techniques to better incorporate audio information in AVSP models. We choose ViNet (being an outperforming model over STAViS) as base model for all our further experiments.

Table 5.1: Comparison of metrics on passing zero and random sound signal. Here [STA-0] and [STA-R] denotes the inference of STA on zero and random sound signal respectively. Similarly [AViNet-0] and [AViNet-R] denotes the inference of AViNet on zero and random sound signal respectively.

	DIEM					*Coutrot1*					*Coutrot2*				
	CC	sAUC	AUC	NSS	SIM	CC	sAUC	AUC	NSS	SIM	CC	sAUC	AUC	NSS	SIM
STAViS(ST)	0.567	0.664	0.879	2.190	0.472	0.459	0.576	0.862	1.990	0.384	0.653	0.689	0.941	4.190	0.447
STAViS(STA)	0.579	0.675	0.883	2.260	0.482	0.472	0.585	0.868	2.110	0.394	0.735	0.710	0.958	5.280	0.511
STAViS(STA-0)	0.576	0.673	0.883	2.249	0.484	0.471	0.584	0.867	2.112	0.396	0.731	0.708	0.956	5.242	0.526
STAViS(STA-R)	0.576	0.673	0.883	2.250	0.484	0.472	0.584	0.867	2.112	0.396	0.731	0.708	0.956	5.233	0.525
ViNet	0.626	0.723	0.898	2.470	0.483	0.551	0.633	0.886	2.680	0.423	0.724	0.739	0.950	5.610	0.466
AViNet	0.632	0.719	0.899	2.530	0.498	0.560	0.635	0.889	2.730	0.425	0.754	0.742	0.951	5.950	0.493
AViNet-0	0.619	0.717	0.897	2.484	0.486	0.558	0.636	0.889	2.727	0.424	0.760	0.748	0.959	6.009	0.494
AViNet-R	0.619	0.717	0.897	2.484	0.486	0.558	0.636	0.889	2.727	0.424	0.760	0.748	0.959	6.010	0.495
	AVAD					*ETMD*					*SumMe*				
	CC	sAUC	AUC	NSS	SIM	CC	sAUC	AUC	NSS	SIM	CC	sAUC	AUC	NSS	SIM
STAViS(ST)	0.604	0.590	0.915	3.070	0.443	0.560	0.727	0.929	2.840	0.412	0.418	0.647	0.884	1.980	0.332
STAViS(STA)	0.608	0.593	0.919	3.180	0.457	0.569	0.731	0.931	2.940	0.425	0.422	0.656	0.888	2.040	0.337
STAViS(STA-0)	0.606	0.592	0.919	3.166	0.463	0.569	0.731	0.931	2.937	0.431	0.422	0.656	0.888	2.038	0.341
STAViS(STA-R)	0.605	0.592	0.919	3.160	0.462	0.569	0.731	0.931	2.936	0.431	0.423	0.656	0.888	2.037	0.340
ViNet	0.694	0.663	0.928	3.820	0.504	0.569	0.736	0.928	3.060	0.409	0.466	0.696	0.898	2.400	0.345
AViNet	0.674	0.658	0.927	3.770	0.491	0.571	0.733	0.928	3.080	0.406	0.463	0.692	0.897	2.410	0.343
AViNet-0	0.673	0.659	0.928	3.759	0.490	0.571	0.733	0.928	3.078	0.407	0.459	0.691	0.896	2.386	0.342
AViNet-R	0.673	0.658	0.928	3.760	0.490	0.570	0.733	0.928	3.074	0.407	0.459	0.692	0.896	2.386	0.342

5.1.2 Analysis of Different Audio Modules

Audio module added to ViNet might not be able to capture contrasting features to video module. We tried some SOTA audio modules that showcased high performance on audio-visual tasks *i.e.*sound source localization[1], active speaker detection[61], audio-visual objects learning[48], etc. Table 5.2 shows the performance on different audio modules. We observe a similar performance across all, thus limiting the learning ability of the network to some extent.

For instance, we observed that AViNet, when equipped with AVID audio module, exhibited the best performance on the AVAD dataset, while SoundNet emerged as the top-performing audio module for the Coutrot2 dataset. Remarkably, this consistent trend of minimal variance in performance among different audio modules was observed across all datasets.

To this end, we adopt different fusion techniques to integrate the audio and visual features in a better way.

Table 5.2: Comparison of metrics on AViNet with different audio modules. Here, [$\text{AViNet}_{\text{SoundNet}}$] , [$\text{AViNet}_{\text{VGG-Vox}}$] and [$\text{AViNet}_{\text{AVID}}$] denotes AViNet with sound encoder as SoundNet, VGG-Vox and AVID respectively.

	DIEM					*Coutrot1*					*Coutrot2*				
	CC	sAUC	AUC	NSS	SIM	CC	sAUC	AUC	NSS	SIM	CC	sAUC	AUC	NSS	SIM
ViNet	0.626	0.723	0.898	2.470	0.483	0.551	0.633	0.886	2.680	0.423	0.724	0.739	0.950	5.61	0.466
$\text{AViNet}_{\text{SoundNet}}$	0.632	0.719	0.899	2.530	0.498	0.560	0.635	0.889	2.730	0.425	0.754	0.742	0.951	5.950	0.493
$\text{AViNet}_{\text{VGG-Vox}}$	0.633	0.732	0.906	2.563	0.494	0.555	0.640	0.891	2.691	0.424	0.749	0.747	0.964	5.829	0.465
$\text{AViNet}_{\text{AVID}}$	0.624	0.722	0.900	2.492	0.488	0.556	0.638	0.890	2.685	0.422	0.721	0.737	0.958	5.653	0.460
	AVAD					*ETMD*					*SumMe*				
	CC	sAUC	AUC	NSS	SIM	CC	sAUC	AUC	NSS	SIM	CC	sAUC	AUC	NSS	SIM
ViNet	0.694	0.663	0.928	3.82	0.504	0.569	0.736	0.928	3.06	0.409	0.466	0.696	0.898	2.40	0.345
$\text{AViNet}_{\text{SoundNet}}$	0.674	0.658	0.927	3.770	0.491	0.571	0.733	0.928	3.080	0.406	0.463	0.692	0.897	2.410	0.343
$\text{AViNet}_{\text{VGG-Vox}}$	0.678	0.660	0.928	2.719	0.488	0.564	0.737	0.928	3.063	0.401	0.462	0.706	0.899	2.382	0.339
$\text{AViNet}_{\text{AVID}}$	0.684	0.662	0.929	3.813	0.494	0.567	0.738	0.928	3.066	0.401	0.462	0.699	0.898	2.384	0.339

5.1.3 Analysis of Different Fusion Techniques

In multi-modal networks, the fusion technique plays a major role. We adopt different fusion techniques that have shown encouraging performance in different multi-modal scenarios. Table 5.3 compares the effect of different fusion techniques on network's performance. For example, our experimental results showed that AViNet, when trained with RNA loss, achieved superior performance on the Coutrot2 dataset, while the Attention-based mechanism network emerged as the top-performing fusion

technique for the ETMD dataset. A minimalistic jitter in results is observed consistently across all datasets, suggesting that different fusion techniques fail to leverage audio in AVSP models.

We believe that one possible reason is that audio information is futile to the video saliency with the existing datasets. Furthermore, the other possible reason could be that the visual network is dominant. This dominance problem might arise because of *norm unbalance* between the two modalities, so that modality with greater feature norm (visual in our case) gets privileged while penalizing the other (audio). To this end, we tried incorporating RNA loss[53] to bring out norm balance and leverage audio in a better way. Table 5.4 shows the norm values before and after applying RNA Loss. The balanced norm suggests that empowering the audio features doesn't benefit the task and visual features are rich enough to predict the final saliency.

Table 5.3: Comparison of metrics on AViNet with different fusion techniques. Here, [AViNet(B)], [AViNet(C)], [AViNet(A)] and [AViNet(RNA)] denotes AViNet with fusion based on Bi-linear, Concatenation, Attention-based mechanism, and RNA loss respectively.

	DIEM					*Coutrot1*					*Coutrot2*				
	CC	sAUC	AUC	NSS	SIM	CC	sAUC	AUC	NSS	SIM	CC	sAUC	AUC	NSS	SIM
ViNet	0.626	0.723	0.898	2.470	0.483	0.551	0.633	0.886	2.680	0.423	0.724	0.739	0.950	5.610	0.466
AViNet(B)	0.632	0.719	0.899	2.530	0.498	0.560	0.635	0.889	2.730	0.425	0.754	0.742	0.951	5.950	0.493
AViNet(C)	0.631	0.720	0.897	2.500	0.497	0.556	0.636	0.887	2.680	0.426	0.753	0.743	0.951	5.810	0.486
AViNet(A)	0.6143	0.707	0.897	2.458	0.488	0.552	0.632	0.890	2.700	0.425	0.744	0.739	0.961	5.776	0.479
AViNet(RNA)	0.621	0.719	0.896	2.470	0.485	0.542	0.624	0.884	2.592	0.413	0.766	0.747	0.961	5.961	0.489
	AVAD					*ETMD*					*SumMe*				
	CC	sAUC	AUC	NSS	SIM	CC	sAUC	AUC	NSS	SIM	CC	sAUC	AUC	NSS	SIM
ViNet	0.694	0.663	0.928	3.820	0.504	0.569	0.736	0.928	3.060	0.409	0.466	0.696	0.898	2.400	0.345
AViNet(B)	0.674	0.658	0.927	3.770	0.491	0.571	0.733	0.928	3.080	0.406	0.463	0.692	0.897	2.410	0.343
AViNet(C)	0.683	0.661	0.931	3.740	0.494	0.566	0.737	0.928	3.050	0.404	0.471	0.699	0.899	2.420	0.346
AViNet(A)	0.674	0.659	0.927	3.726	0.490	0.575	0.735	0.929	3.086	0.413	0.462	0.693	0.897	2.400	0.342
AViNet(RNA)	0.665	0.660	0.928	3.649	0.473	0.565	0.737	0.928	3.032	0.403	0.446	0.686	0.893	2.235	0.331

5.1.4 Why is AV network better than visual-only network?

From above experiments we observe, while audio-visual models achieve outstanding performance compared to visual-only models there still remain an important issue, that is lacking the utilization of audio features. Audio being agnostic, suggest that the AV model somehow empowers the potential capacity of the visual only model. We believe that one possible reason is that the visual only models are not optimal and a regularization technique on Visual model can help to learn the saliency of similar or higher precision.

Table 5.4: Mean and standard deviation of feature norm before and after applying RNA Loss

	AViNet (with Bi-Linear Fusion)		*AViNet with RNA Loss*	
	Audio	Video	Audio	Video
AVAD	9.5142 ± 4.7232	29.0128 ± 3.9406	11.8473 ± 4.3091	14.2908 ± 2.6662
Coutrot1	9.3178 ± 4.9157	25.9076 ± 3.3296	11.6736 ± 4.4389	11.9309 ± 2.1535
Coutrot2	13.5336 ± 1.8181	26.6241 ± 1.3176	15.136 ± 2.1201	12.5556 ± 1.2305
DIEM	11.4565 ± 3.8720	28.3217 ± 5.4160	13.3933 ± 3.8439	12.5178 ± 2.1789
ETMD	11.3443 ± 4.1168	27.4783 ± 4.1482	13.4269 ± 3.4984	12.9234 ± 1.712
SumMe	10.0412 ± 4.6872	27.1688 ± 4.6831	12.7507 ± 4.4734	12.5161 ± 2.5662

5.1.5 Regularization of visual features

5.1.5.1 Random Blacking of Frames

To regularize the visual-only model based on temporal information, we tried a simple technique by random blacking of frames, i.e., ViNet takes 32 frames as input. Here we randomly blacked 4,8,16 frames respectively. The visual-only network is forced to predict the saliency of the last frame by the input as the last 32 frames. Thus the information on the saliency of last frames is highly biased on some last frames as compared to the initial frames, there is a high chance of the network ignoring the importance of all the temporal information and relying on a specific section of the frames. Here blacking of frames validates the importance of each frame, as there might be a scenario where the last frame information is being blacked. We tried experiments by different ablation studies on varying the number of frames being randomly blacked: 4,8,16.

Table 5.5 shows the performance on applying different blackening schemes. Here, the minimalistic performance drop in ViNet-4, ViNet-8, ViNet-16 respectively validates the temporal information utilization of ViNet.Since, ViNet-4, ViNet-8 doesn't show any significant performance gain we can claim that ViNet itself is utilizing the temporal information and weighing each sequential frame with some importance. The incorporation of regularization by random blacking of frames did not adversely impact the performance of our model, nor did it compromise our primary objective of utilizing visual features in a more optimal manner. Notably, consistent performance was observed across all datasets, including DIEM, where the performance remained unchanged.

5.1.5.2 Regularization by Vanilla DropOut

In order to investigate the impact of audio on the fusion of audio-visual features, we analyzed the output feature vectors from the audio module. We selected a single batch from our audio-visual dataset and found that 77% (794 out of 1024 audio features)of the audio features were zero-valued. This

Table 5.5: Comparison of metrics on the *DIEM*, *Coutrot1*, *Coutrot2*, *AVAD*, *ETMD* and *SumMe* test sets. Here, ViNet-4, ViNet-8, and ViNet-16 refer to the model trained by random blacking of 4,8, and 16 frames, respectively.

	DIEM					*Coutrot1*					*Coutrot2*				
	CC	sAUC	AUC	NSS	SIM	CC	sAUC	AUC	NSS	SIM	CC	sAUC	AUC	NSS	SIM
ViNet	0.626	0.723	0.898	2.470	0.483	0.551	0.633	0.886	2.680	0.423	0.724	0.739	0.950	5.610	0.466
AViNet	0.632	0.719	0.899	2.530	0.498	0.560	0.635	0.889	2.730	0.425	0.754	0.742	0.951	5.950	0.493
ViNet-4	0.626	0.722	0.898	2.47	0.484	0.55	0.632	0.886	2.68	0.423	0.723	0.739	0.949	5.61	0.465
ViNet-8	0.625	0.723	0.897	2.47	0.483	0.549	0.632	0.885	2.68	0.422	0.723	0.738	0.949	5.6	0.464
ViNet-16	0.622	0.72	0.892	2.45	0.479	0.537	0.611	0.878	2.62	0.412	0.701	0.712	0.942	5.48	0.443
	AVAD					*ETMD*					*SumMe*				
	CC	sAUC	AUC	NSS	SIM	CC	sAUC	AUC	NSS	SIM	CC	sAUC	AUC	NSS	SIM
ViNet	0.694	0.663	0.928	3.820	0.504	0.569	0.736	0.928	3.060	0.409	0.466	0.696	0.898	2.400	0.345
AViNet	0.674	0.658	0.927	3.770	0.491	0.571	0.733	0.928	3.080	0.406	0.463	0.692	0.897	2.410	0.343
ViNet-4	0.693	0.662	0.928	3.81	0.504	0.568	0.734	0.925	3.05	0.408	0.465	0.696	0.897	2.4	0.344
ViNet-8	0.692	0.662	0.927	3.81	0.502	0.567	0.734	0.924	3.04	0.408	0.464	0.695	0.896	2.38	0.342
ViNet-16	0.681	0.621	0.911	3.74	0.472	0.523	0.706	0.904	2.99	0.396	0.438	0.643	0.838	2.19	0.305

surprising observation suggests that during the fusion of audio and visual features, these zero-valued audio features may nullify the visual features to a significant extent. This behavior is reminiscent of the well-known technique called Dropout, which aims to regularize the model by randomly setting certain feature values to zero during training. Based on this observation, we applied a variant of Dropout, known as Vanilla Dropout, on our ViNet model across all datasets to further investigate and potentially mitigate this behavior.

Vanilla DropOut is conventionally used to regularize deep CNNs. Table 5.6 illustrates the results on varying dropout and using dropout of 0.85 gave better results on which all our further analysis is carried out. The comparison of visual and audio visual models with regularized visual model are presented in Table 5.7. The regularized model is able to recover most of the underlying performance on current datasets. The results shows a similar behaviour in regularized model and the audio-visual model with respect to the visual only model. On specific dataset like Coutrot2, where the audio visual model seemed to gain significant improvement, our results indicates the similar significant gain by the regularized model. Thus audio visual model can be surmised as some form of regularization applied over visual only model.

Table 5.6: Results on varying Dropout on Coutrot2 test set.

	STAViS					*ViNet*				
Dropout	CC	sAUC	AUC	NSS	SIM	CC	sAUC	AUC	NSS	SIM
0.80	0.674	0.715	0.952	4.427	0.308	0.735	0.740	0.955	5.761	0.481
0.85	0.675	0.715	0.955	4.432	0.309	0.740	0.741	0.959	5.777	0.481
0.90	0.673	0.713	0.948	4.397	0.294	0.733	0.739	0.954	5.748	0.481

Table 5.7: Comparison of metrics on the *DIEM*, *Coutrot1*, *Coutrot2*, *AVAD*, *ETMD* and *SumMe* test sets. Here, STAViS(STD) and ViNet-D refers to respective regularized models with 85% dropout

	DIEM					*Coutrot1*					*Coutrot2*				
	CC	sAUC	AUC	NSS	SIM	CC	sAUC	AUC	NSS	SIM	CC	sAUC	AUC	NSS	SIM
STAViS(ST)	0.566	0.664	0.879	2.190	0.471	0.458	0.576	0.861	1.990	0.384	0.653	0.689	0.940	4.190	0.447
STAViS(STA)	0.579	0.674	0.883	2.260	0.482	0.472	0.584	0.868	2.110	0.393	0.735	0.710	0.958	5.280	0.511
STAViS(STD)	0.609	0.693	0.890	2.329	0.406	0.509	0.593	0.876	2.202	0.338	0.675	0.714	0.955	4.432	0.309
ViNet	0.626	0.723	0.898	2.470	0.483	0.551	0.633	0.886	2.680	0.423	0.724	0.739	0.950	5.610	0.466
AViNet	0.632	0.719	0.899	2.530	0.498	0.560	0.635	0.889	2.730	0.425	0.754	0.742	0.951	5.950	0.493
ViNet-D	0.637	0.724	0.902	2.559	0.498	0.561	0.634	0.891	2.736	0.430	0.740	0.741	0.959	5.777	0.481
	AVAD					*ETMD*					*SumMe*				
	CC	sAUC	AUC	NSS	SIM	CC	sAUC	AUC	NSS	SIM	CC	sAUC	AUC	NSS	SIM
STAViS(ST)	0.604	0.59	0.915	3.070	0.443	0.560	0.727	0.929	2.840	0.412	0.418	0.647	0.884	1.980	0.332
STAViS(STA)	0.608	0.593	0.919	3.180	0.457	0.569	0.731	0.931	2.940	0.425	0.422	0.656	0.888	2.040	0.337
STAViS(STD)	0.609	0.600	0.919	3.078	0.345	0.562	0.744	0.929	2.835	0.314	0.443	0.676	0.893	2.135	0.274
ViNet	0.694	0.663	0.928	3.820	0.504	0.569	0.736	0.928	3.060	0.409	0.466	0.696	0.898	2.400	0.345
AViNet	0.674	0.658	0.927	3.770	0.491	0.571	0.733	0.928	3.080	0.406	0.463	0.692	0.897	2.410	0.343
ViNet-D	0.682	0.661	0.929	3.835	0.497	0.578	0.740	0.930	3.128	0.416	0.467	0.700	0.899	2.425	0.347

5.1.6 Validation of our hypothesis

As Vanilla Dropout (in Section 5.1.5.2) was able to recover the gains of the Audiovisual model over the visual-only model, we hypothesize that the audio module acts as a regularizer. To this end, we tried a similar random audio regularization technique over different datasets. This random audio was hand-crafted by three different techniques:

- Random vector: generated from normal distribution.

Table 5.8: Results of all the experiments discussed, on a recently proposed large-scale multi-face saliency dataset - MVVA.

	MVVA				
	CC	SIM	NSS	AUC	KLDiv
AViNet(B)	0.7953	0.6006	3.5085	0.8855	0.7582
AViNet-0	0.7962	0.6005	3.5125	0.8856	0.7576
AViNet-R	0.7962	0.6007	3.5125	0.8856	0.7573
AViNet(A)	0.7919	0.5971	3.4919	0.8871	0.7666
AViNet(RNA)	0.7967	0.5991	3.5135	0.8898	0.7603
ViNet-D	0.7956	0.6047	3.5104	0.8849	0.7632
AViNet$_{VGG}$	0.7927	0.6003	3.4912	0.8834	0.7534
AViNet$_{Avid}$	0.7932	0.6034	3.4923	0.8848	0.7573

- Using Wolfram Algorithm: For mimicking the audio distribution of respective video clips, we applied the wolfram algorithm to create random audio based on the distribution of original audio vector.
- Shuffled audio from different clips: Audio from different videos are shuffled and passed to the network in form of random audio (noise or irrelevant audio).

As shown in Table 5.9, the performance of AViNet trained with random audio was found to be similar to AViNet trained with actual audio, supporting the hypothesis that AViNet fails to effectively leverage audio information and instead acts as a regularizer for the visual-only model. These findings provide further evidence of the potential regularization effect of the audio module in AVSP models.

5.2 Multi-Face dataset (MVVA)

To validate and reinforce our findings, we conducted a rigorous set of experiments on the recently proposed MVVA dataset, employing AViNet with the identical settings as described in the previous sections. Notably, we observed that the performance of AViNet, when inferred with zero and random audio inputs, remained similar to that with actual audio inputs, thereby substantiating the negligible impact of audio information on audio-visual saliency prediction models. Subsequently, we performed comprehensive experiments with diverse audio modules and fusion techniques, as discussed in Section 5.1.2 and 5.1.3, and found that the observations were consistent with our previous findings on other datasets. No significant changes in performance were observed, further indicating that the choice of audio module or fusion technique had minimal impact on the overall performance of AViNet on the MVVA dataset. These results further consolidate our hypothesis and emphasize the limited significance

Table 5.9: Comparison of metrics on the *DIEM*, *Coutrot1*, *Coutrot2*, *AVAD*, *ETMD* and *SumMe* test sets. Here, $\text{ViNet}_{\text{random}}$, $\text{ViNet}_{\text{Wolfram}}$ and $\text{ViNet}_{\text{shuffled}}$ refer to models trained on audio vector generated from a normal distribution, Wolfrom algorithm and shuffled audio respectively.

	DIEM					*Coutrot1*					*Coutrot2*				
	CC	sAUC	AUC	NSS	SIM	CC	sAUC	AUC	NSS	SIM	CC	sAUC	AUC	NSS	SIM
ViNet	0.626	0.723	0.898	2.470	0.483	0.551	0.633	0.886	2.680	0.423	0.724	0.739	0.950	5.610	0.466
AViNet-B	0.632	0.719	0.899	2.530	0.498	0.560	0.635	0.889	2.730	0.425	0.754	0.742	0.951	5.950	0.493
AViNet-C	0.631	0.72	0.897	2.5	0.497	0.556	0.636	0.887	2.68	0.426	0.753	0.743	0.951	5.81	0.486
$\text{ViNet}_{\text{random}}$	0.63	0.719	0.898	2.51	0.498	0.557	0.635	0.888	2.71	0.425	0.754	0.741	0.951	5.94	0.492
$\text{ViNet}_{\text{Wolfram}}$	0.632	0.719	0.898	2.52	0.498	0.559	0.636	0.889	2.72	0.426	0.753	0.741	0.95	5.94	0.492
$\text{ViNet}_{\text{shuffled}}$	0.631	0.719	0.898	2.53	0.498	0.558	0.635	0.889	2.72	0.426	0.754	0.742	0.951	5.94	0.492
	AVAD					*ETMD*					*SumMe*				
	CC	sAUC	AUC	NSS	SIM	CC	sAUC	AUC	NSS	SIM	CC	sAUC	AUC	NSS	SIM
ViNet	0.694	0.663	0.928	3.820	0.504	0.569	0.736	0.928	3.060	0.409	0.466	0.696	0.898	2.400	0.345
AViNet-B	0.674	0.658	0.927	3.770	0.491	0.571	0.733	0.928	3.080	0.406	0.463	0.692	0.897	2.410	0.343
AViNet-C	0.683	0.661	0.931	3.74	0.494	0.566	0.737	0.928	3.05	0.404	0.471	0.699	0.899	2.42	0.346
$\text{ViNet}_{\text{random}}$	0.673	0.659	0.928	3.77	0.491	0.57	0.732	0.927	3.07	0.405	0.462	0.69	0.896	2.4	0.342
$\text{ViNet}_{\text{Wolfram}}$	0.674	0.66	0.928	3.77	0.492	0.571	0.731	0.928	3.08	0.405	0.461	0.691	0.896	2.4	0.342
$\text{ViNet}_{\text{shuffled}}$	0.673	0.659	0.927	3.77	0.49	0.57	0.733	0.927	3.07	0.406	0.462	0.691	0.896	2.41	0.342

of audio information in the context of audio-visual saliency prediction models. Intriguingly, we observed that dropout regularization appeared to inhibit the incremental gains in performance from ViNet to AViNet (in Table 5.8). These results further validate our observations and provide additional evidence of the impact of audio-visual integration and dropout regularization on the performance of AVSP models.

5.3 Visual Only Datasets

Our hypothesis was further validated through experiments conducted on visual-only datasets, where the audio input was set to a random vector. The results, as shown in Table 5.10, clearly demonstrate that AViNet exhibits performance that is similar to, or even better than, ViNet. Notably, on the UCF dataset, a significant improvement can be observed, with the correlation coefficient (CC) increasing from 0.673 to 0.723, leading to state-of-the-art (SOTA) performance by simply adding a random audio module. Furthermore, it is worth noting that the dropout regularization technique inhibits this behavior, suggesting that the previously claimed audio-visual models may not effectively incorporate audio information, but rather regularize the visual module.

These findings highlight the need for careful consideration of the role of audio information in audio-visual models, and suggest that previously claimed audio-visual models may primarily function as regularizers for the visual module. Overall, our study provides insights into the nuanced interplay between audio and visual modalities in audio-visual models, and contributes to a better understanding of the underlying mechanisms and performance dynamics in this field.

Table 5.10: Comparison of metrics on the *DHF1K(val)*, *Hollywood-2* and *UCF-Sports* test sets.

	DHF1K					*Hollywood-2*					*UCF-Sports*				
	CC	sAUC	AUC	NSS	SIM	CC	sAUC	AUC	NSS	SIM	CC	sAUC	AUC	NSS	SIM
ViNet	0.521	0.732	0.919	2.956	0.388	0.693	0.813	0.930	3.730	0.550	0.673	0.810	0.924	3.620	0.522
AViNet	0.517	0.723	0.912	2.941	0.380	0.700	0.814	0.931	3.661	0.534	0.709	0.809	0.931	3.915	0.531
ViNet-D	0.521	0.729	0.914	3.000	0.379	0.703	0.815	0.930	3.778	0.551	0.723	0.812	0.936	3.956	0.533

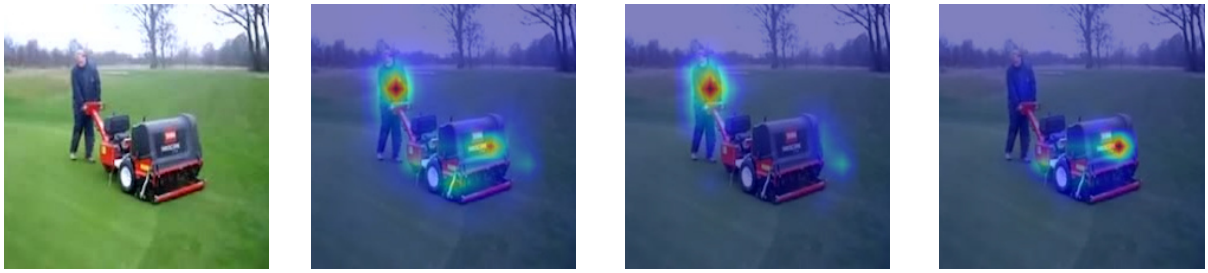
Chapter 6

Conclusions and Future Work

This thesis presents a comprehensive analysis to underline the role of audio in current deep AVSP methods. Our experiments on 10 different datasets clearly indicate that visual modality dominates the learning; the current models largely ignore the audio information. The observation is consistent while using three different audio backbones and four different fusion techniques. The observations contrast with the previous methods, which claim audio as a significant contributing factor. We show the performance gains are a byproduct of improved training and the additional audio branch seems to have a regularizing effect. We show that similar gains are achieved while sending random audio during training.

The results demonstrate a clear gap between human learning and deep learning-based models. Several psycho-visual studies show that audio impacts visual attention; however, neural networks seem to discard this information. We believe there could be multiple reasons behind the finding. First, neural networks behave differently than humans. For instance, in a multi-person conversation, humans exhibit turn-taking behavior. In contrast, networks can process all faces (or lip movements) in parallel through the convolution filters.

Limitations of the dataset could be the second reason for this. For instance, if the actions are highly correlated with sound, localizing movement/actions can help predict saliency. Similarly, in datasets with frontal face conversations, just picking the lip movement can help identify the speaker and aid saliency prediction, and audio modality might be ignored. Finally, one major limitation of all works discussed in the paper is that they use monaural audio, and hence the directional aspect is discarded. In contrast, the ability of humans to sense the direction of the audio significantly aids the attention mechanism. A future direction [13] could be to curate large-scale datasets with directional audio (stereo) and 360-degree videos. The monaural audio and limited field of view can then be simulated from such datasets.



(a) Original Image (b) Potential Loc. map (c) Human sound (d) Machine sound

Figure 6.1: Visualization of localization maps with several objects capable of producing sound in the image and responding to an object that is n producing sound.

Also to effectively incorporate audio cues, there is a need to curate datasets in the direction of Sound Source Localization [50] which falls in line with human attention and tends to fixate it in the surroundings by classifying audio to different objects. For example, as shown in the Figure 6.1, the localization map of machine sound (Fig. 6.1d) tends to fixate on the machine, whereas the localization map of a human's sound fixates on the person operating the machine (Fig. 6.1c).

Overall, we believe the experiments presented in this thesis will help the community reflect upon the role of audio in the current research landscape, identify the shortcomings, and help build improved AVSP models.

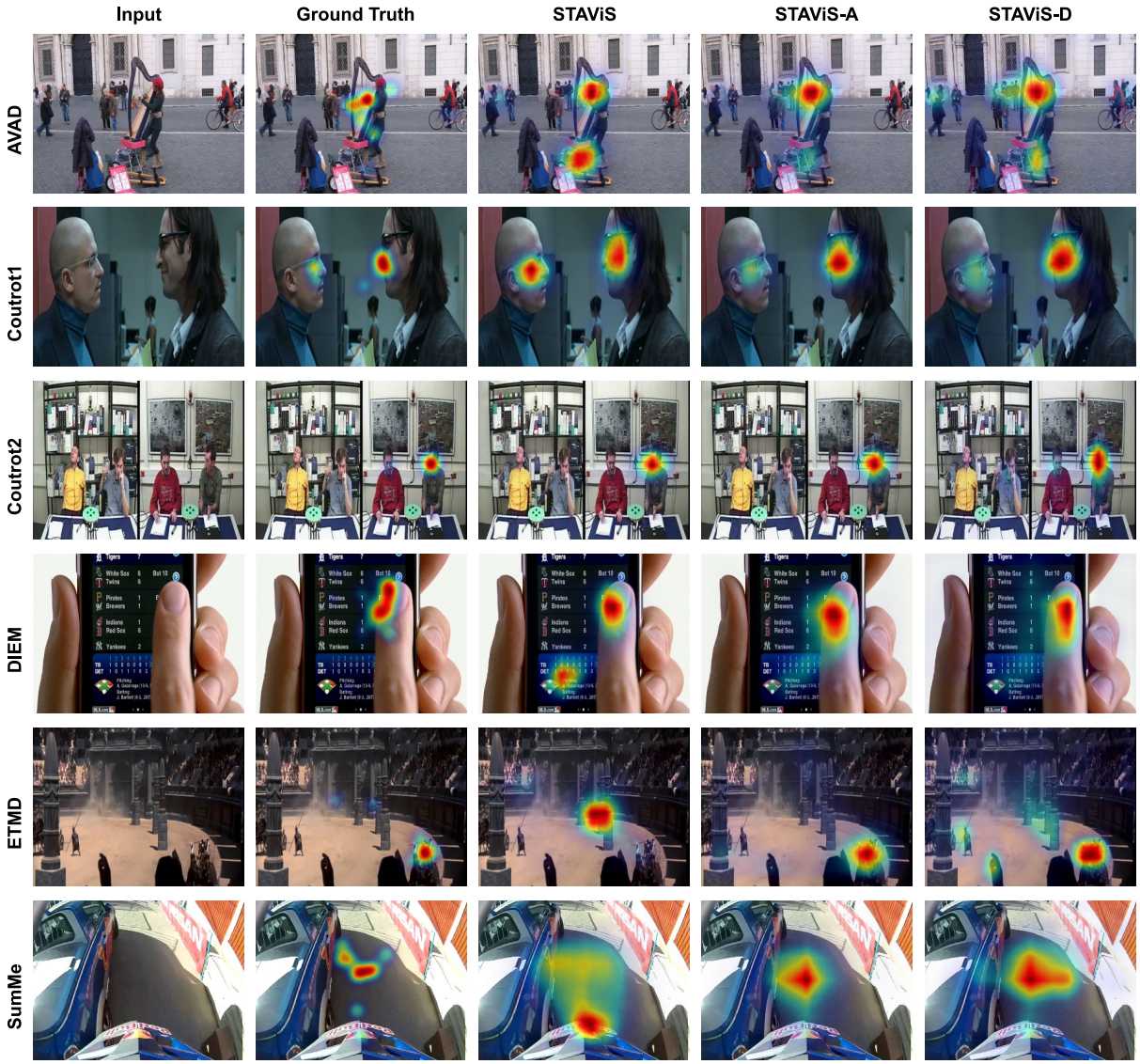


Figure 6.2: Qualitative Comparisons of our hypothesis on 6 Saliency Datasets for STAViS (Here dropout is chosen as 85%)

Figure 6.3: Qualitative Comparisons of our hypothesis on 10 Saliency Datasets for AViNet (Here dropout is chosen as 85%)

List of Publications

- **Ritvik Agrawal** [*] , Shreyank Jyoti [*], Rohit Girmaji, Sarath Sivaprasad, Vineet Gandhi. *"Does Audio help in deep Audio-Visual Saliency prediction models?"*. In the **ACM International Conference on Multimodal Interaction (ICMI)** 2022 [ORAL](Outstanding Student Paper Award).
- Shreyank Jyoti [*], Rohit Girmaji[*] ,**Ritvik Agrawal** [*], Sarath Sivaprasad, Vineet Gandhi. *"Salient Face Prediction without Bells and Whistles"*. In the **IEEE International Conference on Digital Image Computing: Techniques and Applications (DICTA)** 2022 [ORAL]. (Not part of thesis)

[*] denotes Equal Contribution

Bibliography

[1] Y. Aytar, C. Vondrick, and A. Torralba. Soundnet: Learning sound representations from unlabeled video. *Advances in neural information processing systems*, 29, 2016. 12, 13, 28

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

[3] G. Bellitto, F. Proietto Salanitri, S. Palazzo, F. Rundo, D. Giordano, and C. Spampinato. Hierarchical domain-adapted feature learning for video saliency prediction. *International Journal of Computer Vision*, 129(12):3216–3232, 2021. 9

[4] A. Borji and L. Itti. Scene classification with a sparse set of salient regions. In *2011 IEEE International Conference on Robotics and Automation*, pages 1902–1908. IEEE, 2011. 2

[5] N. J. Butko, L. Zhang, G. W. Cottrell, and J. R. Movellan. Visual saliency model for robot cameras. In *2008 IEEE International Conference on Robotics and Automation*, pages 2398–2403. IEEE, 2008. 2

[6] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand. What do different evaluation metrics tell us about saliency models? *IEEE transactions on pattern analysis and machine intelligence*, 41(3):740–757, 2018. 22

[7] J. Carreira and A. 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. 11, 12

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

[9] J. Chen, Q. Li, H. Ling, D. Ren, and P. Duan. Audiovisual saliency prediction via deep learning. *Neurocomputing*, 428:248–258, 2021. 10

[10] L. Chen, P. Huang, and Z. Zhao. Saliency based proposal refinement in robotic vision. In *2017 IEEE International Conference on Real-time Computing and Robotics (RCAR)*, pages 85–90. IEEE, 2017. 2

[11] Y. Chen, T. V. Nguyen, M. Kankanhalli, J. Yuan, S. Yan, and M. Wang. Audio matters in visual attention. *IEEE Transactions on Circuits and Systems for Video Technology*, 24(11):1992–2003, 2014. 6

[12] J. S. Chung, A. Nagrani, and A. Zisserman. VoxCeleb2: Deep Speaker Recognition. In *Proc. Interspeech 2018*, pages 1086–1090, 2018. 13

[13] M. Cokelek, N. Imamoglu, C. Ozcinar, E. Erdem, and A. Erdem. Leveraging frequency based salient spatial sound localization to improve 360° video saliency prediction. In *2021 17th International Conference on Machine Vision and Applications (MVA)*, pages 1–5. IEEE, 2021. 36

[14] A. Coutrot and N. Guyader. How saliency, faces, and sound influence gaze in dynamic social scenes. *Journal of vision*, 14(8):5–5, 2014. 3, 7, 21

[15] A. Coutrot and N. Guyader. An efficient audiovisual saliency model to predict eye positions when looking at conversations. In *2015 23rd European Signal Processing Conference (EUSIPCO)*, pages 1531–1535. IEEE, 2015. 7, 10

[16] A. Coutrot and N. Guyader. Multimodal saliency models for videos. In *From Human Attention to Computational Attention*. 2016. 3

[17] A. Coutrot, N. Guyader, G. Ionescu, and A. Caplier. Influence of soundtrack on eye movements during video exploration. *Journal of Eye Movement Research*, 5(4):2, 2012. 2, 7

[18] A. Coutrot, N. Guyader, G. Ionescu, and A. Caplier. Video viewing: do auditory salient events capture visual attention? *annals of telecommunications-annales des télécommunications*, 69(1):89–97, 2014. 7

[19] R. Droste, J. Jiao, and J. A. Noble. Unified image and video saliency modeling. In *European Conference on Computer Vision*, pages 419–435. Springer, 2020. 8

[20] J. F. Ferreira and J. Dias. Attentional mechanisms for socially interactive robots–a survey. *IEEE Transactions on Autonomous Mental Development*, 6(2):110–125, 2014. 2

[21] S. Frintrop and M. Kessel. Most salient region tracking. In *2009 IEEE International Conference on Robotics and Automation*, pages 1869–1874. IEEE, 2009. 2

[22] W. W. Gaver. What in the world do we hear?: An ecological approach to auditory event perception. *Ecological psychology*, 5(1):1–29, 1993. 3

[23] S. Gorji and J. J. Clark. Going from image to video saliency: Augmenting image salience with dynamic attentional push. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7501–7511, 2018. 8

[24] M. Gygli, H. Grabner, H. Riemenschneider, and L. V. Gool. Creating summaries from user videos. In *European conference on computer vision*, pages 505–520. Springer, 2014. 3, 21

[25] H. Hadizadeh and I. V. Bajić. Saliency-aware video compression. *IEEE Transactions on Image Processing*, 23(1):19–33, 2013. 2

[26] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 20(11):1254–1259, 1998. 2

[27] S. Jain, P. Yarlagadda, S. Jyoti, S. Karthik, R. Subramanian, and V. Gandhi. Vinet: Pushing the limits of visual modality for audio-visual saliency prediction. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3520–3527. IEEE, 2020. 9, 10, 11, 12, 15, 22

[28] L. Jiang, M. Xu, T. Liu, M. Qiao, and Z. Wang. Deepvs: A deep learning based video saliency prediction approach. In *Proceedings of the european conference on computer vision (eccv)*, pages 602–617, 2018. 8

[29] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 3

[30] P. Khuwuthyakorn, A. Robles-Kelly, and J. Zhou. Object of interest detection by saliency learning. In *European conference on Computer vision*, pages 636–649. Springer, 2010. 2

[31] P. Koutras and P. Maragos. A perceptually based spatio-temporal computational framework for visual saliency estimation. *Signal Processing: Image Communication*, 38:15–31, 2015. 3, 22

[32] P. Koutras and P. Maragos. Susinet: See, understand and summarize it. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 3, 10, 11

[33] Q. Lai, W. Wang, H. Sun, and J. Shen. Video saliency prediction using spatiotemporal residual attentive networks. *IEEE Transactions on Image Processing*, 29:1113–1126, 2019. 8

[34] P. Linardos, E. Mohedano, J. J. Nieto, N. E. O'Connor, X. Giro-i Nieto, and K. McGuinness. Simple vs complex temporal recurrences for video saliency prediction. *arXiv preprint arXiv:1907.01869*, 2019. 8

[35] Y. Liu, M. Qiao, M. Xu, B. Li, W. Hu, and A. Borji. Learning to predict salient faces: A novel visual-audio saliency model. In *European Conference on Computer Vision*, pages 413–429. Springer, 2020. 3, 10, 22

[36] Y. Liu, S. Zhang, M. Xu, and X. He. Predicting salient face in multiple-face videos. In *CVPR*, 2017. 12

[37] D. Man and R. Olchawa. Brain biophysics: perception, consciousness, creativity. brain computer interface (bci). In *International Scientific Conference BCI 2018 Opole*, pages 38–44. Springer, 2018. 6

[38] M. Mancas, V. P. Ferrera, N. Riche, and J. G. Taylor. *From Human Attention to Computational Attention*, volume 2. Springer, 2016. 3, 21

[39] P. Marighetto, A. Coutrot, N. Riche, N. Guyader, M. Mancas, B. Gosselin, and R. Laganiere. Audio-visual attention: Eye-tracking dataset and analysis toolbox. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 1802–1806. IEEE, 2017. 7

[40] M. Marszalek, I. Laptev, and C. Schmid. Actions in context. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2929–2936. IEEE, 2009. 3, 20

[41] H. McGurk and J. MacDonald. Hearing lips and seeing voices. *Nature*, 1976. 7

[42] K. Min and J. 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. 9

[43] X. Min, G. Zhai, Z. Gao, C. Hu, and X. Yang. Sound influences visual attention discriminately in videos. In *2014 Sixth International Workshop on Quality of Multimedia Experience (QoMEX)*, pages 153–158. IEEE, 2014. 2

[44] X. Min, G. Zhai, K. Gu, and X. Yang. Fixation prediction through multimodal analysis. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 13(1):1–23, 2016. 3, 21

[45] X. Min, G. Zhai, J. Zhou, X.-P. Zhang, X. Yang, and X. Guan. A multimodal saliency model for videos with high audio-visual correspondence. *IEEE Transactions on Image Processing*, 29:3805–3819, 2020. 3

[46] P. K. Mital, T. J. Smith, R. L. Hill, and J. M. Henderson. Clustering of gaze during dynamic scene viewing is predicted by motion. *Cognitive computation*, 3(1):5–24, 2011. 3, 21

[47] K. B. Moorthy, M. Kumar, R. Subramanian, and V. Gandhi. Gazed–gaze-guided cinematic editing of wide-angle monocular video recordings. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–11, 2020. 2

[48] P. Morgado, N. Vasconcelos, and I. Misra. Audio-visual instance discrimination with cross-modal agreement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12475–12486, June 2021. 13, 14, 28

[49] T. V. Nguyen, M. Xu, G. Gao, M. Kankanhalli, Q. Tian, and S. Yan. Static saliency vs. dynamic saliency: a comparative study. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 987–996, 2013. 2

[50] T. Oya, S. Iwase, R. Natsume, T. Itazuri, S. Yamaguchi, and S. Morishima. Do we need sound for sound source localization? In *Proceedings of the Asian Conference on Computer Vision*, 2020. 37

[51] M. S. Pápai and S. Soto-Faraco. Sounds can boost the awareness of visual events through attention without cross-modal integration. *Scientific reports*, 7(1):1–13, 2017. 6

[52] D. R. Perrott, K. Saberi, K. Brown, and T. Z. Strybel. Auditory psychomotor coordination and visual search performance. *Perception & psychophysics*, 48(3):214–226, 1990. 2, 3

[53] M. Planamente, C. Plizzari, E. Alberti, and B. Caputo. Domain generalization through audio-visual relative norm alignment in first person action recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1807–1818, 2022. 17, 18, 29

[54] M. Qiao, Y. Liu, M. Xu, X. Deng, B. Li, W. Hu, and A. Borji. Joint learning of visual-audio saliency prediction and sound source localization on multi-face videos. *arXiv preprint arXiv:2111.08567*, 2021. 3, 10

[55] M. D. Rodriguez, J. Ahmed, and M. Shah. Action mach a spatio-temporal maximum average correlation height filter for action recognition. In *2008 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2008. 3, 21

[56] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 9

[57] V. Sitzmann, A. Serrano, A. Pavel, M. Agrawala, D. Gutierrez, B. Masia, and G. Wetzstein. Saliency in vr: How do people explore virtual environments? *IEEE transactions on visualization and computer graphics*, 24(4):1633–1642, 2018. 2

[58] G. Song, D. Pellerin, and L. Granjon. Sound effect on visual gaze when looking at videos. In *2011 19th European Signal Processing Conference*, pages 2034–2038. IEEE, 2011. 7

[59] G. Song, D. Pellerin, and L. Granjon. Different types of sounds influence gaze differently in videos. *Journal of Eye Movement Research*, 6(4), 2013. 3

[60] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014. 18

[61] R. Tao, Z. Pan, R. K. Das, X. Qian, M. Z. Shou, and H. Li. Is someone speaking? exploring long-term temporal features for audio-visual active speaker detection. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3927–3935, 2021. 16, 28

[62] H. R. Tavakoli, A. Borji, E. Rahtu, and J. Kannala. Dave: A deep audio-visual embedding for dynamic saliency prediction. *arXiv preprint arXiv:1905.10693*, 2019. 3, 10

[63] Y. Tian, J. Shi, B. Li, Z. Duan, and C. Xu. Audio-visual event localization in unconstrained videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 247–263, 2018. 3

[64] A. Tsiami, A. Katsamanis, P. Maragos, and A. Vatakis. Towards a behaviorally-validated computational audiovisual saliency model. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2847–2851. IEEE, 2016. 7

[65] A. Tsiami, P. Koutras, and P. Maragos. Stavis: Spatio-temporal audiovisual saliency network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3, 10, 11, 12, 15, 22

[66] E. Van der Burg, C. N. Olivers, A. W. Bronkhorst, and J. Theeuwes. Pip and pop: nonspatial auditory signals improve spatial visual search. *Journal of Experimental Psychology: Human Perception and Performance*, 34(5):1053, 2008. 7

[67] J. Vroomen and B. d. Gelder. Sound enhances visual perception: cross-modal effects of auditory organization on vision. *Journal of experimental psychology: Human perception and performance*, 26(5):1583, 2000. 2, 3

[68] W. Wang, J. Shen, F. Guo, M.-M. Cheng, and A. 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. 2, 3, 8, 20

[69] P. Warden. Speech commands: A dataset for limited-vocabulary speech recognition. *arXiv preprint arXiv:1804.03209*, 2018. 3

[70] X. Wu, Z. Wu, J. Zhang, L. Ju, and S. Wang. Salsac: A video saliency prediction model with shuffled attentions and correlation-based convlstm. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12410–12417, 2020. 8, 9

[71] S. Xie, C. Sun, J. Huang, Z. Tu, and K. 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. 12

[72] T. Yubing, F. A. Cheikh, F. F. E. Guraya, H. Konik, and A. Trémeau. A spatiotemporal saliency model for video surveillance. *Cognitive Computation*, 3(1):241–263, 2011. 2

[73] K. Zhang, Z. Chen, and S. Liu. A spatial-temporal recurrent neural network for video saliency prediction. *IEEE Transactions on Image Processing*, 30:572–587, 2020. 8

[74] D. Zhu, D. Zhao, X. Min, T. Han, Q. Zhou, S. Yu, Y. Chen, G. Zhai, and X. Yang. Lavs: A lightweight audio-visual saliency prediction model. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2021. 3, 10