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

Ritvik Agrawal*†
Shreyank Jyoti*†
Rohit Girmaji†
Sarath Sivaprasad†‡
Vineet Gandhi†
ritvik.agrawal@research.iiit.ac.in

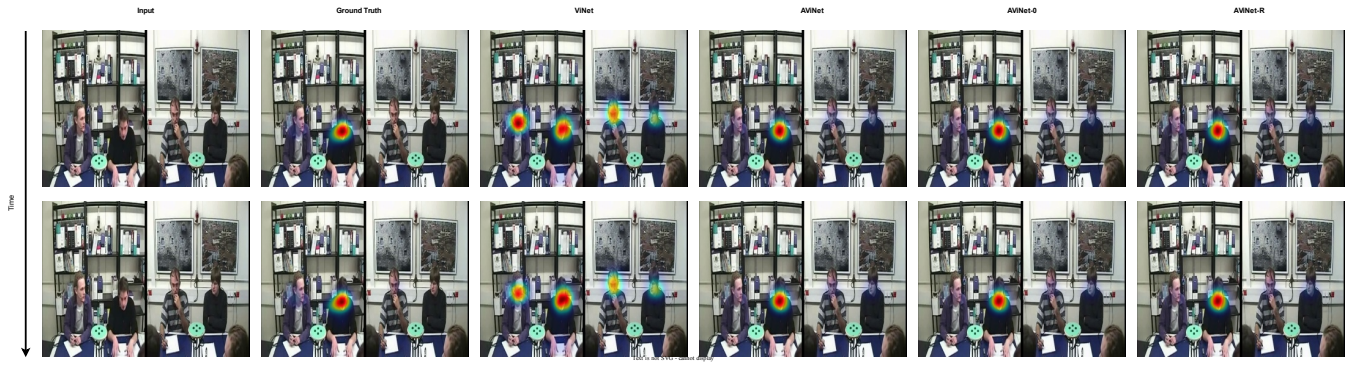

Figure 1: Example frames from a multi-person conversation video (first column) and the corresponding saliency (second column). ViNet is a visual-only saliency prediction model, and AViNet is an audio-visual saliency prediction model. AViNet gives better predictions in this example. On the first reflection, 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 Clearly, the audio is obsolete at inference.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.

ABSTRACT

Despite existing works of Audio-Visual Saliency Prediction (AVSP) models claiming to achieve promising results by fusing audio modality over visual-only models, these models fail to leverage audio information. In this paper, we investigate the relevance of audio cues in conjunction with the visual ones and conduct extensive analysis 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. 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 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.

CCS CONCEPTS

• **Multi-Modal Learning** → **Saliency Prediction**; • **Human Visual System** → *Attention Mechanism.*

KEYWORDS

Human Visual Attention; Saliency Prediction; Multi-modal Learning

ACM Reference Format:
Ritvik Agrawal, Shreyank Jyoti, Rohit Girmaji, Sarath Sivaprasad, and Vineet Gandhi. 2022. Does Audio help in deep Audio-Visual Saliency prediction models?. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '22), November 7–11, 2022, Bengaluru, India.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3536221.3556625

1 INTRODUCTION

The human visual attention (HVA) mechanism facilitates diverse information processing in our surroundings by localizing the most prominent (salient) region[17]. This fundamental ability empowers primates to rapidly analyze/interpret the complex surroundings

*Both authors contributed equally to this research.
†CVIT, KCIS, International Institute for Information Technology, Hyderabad
‡TCS Research, Pune

ICMI '22, November 7–11, 2022, Bengaluru, India

ACM ISBN 978-1-4503-9390-4/22/11…$15.00
https://doi.org/10.1145/3536221.3556625

by locating and devoting the focus only on sub-regions of interest [19]. Mimicking this ability in machines is a fundamental research problem [2] 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 [54], cinematic editing [32], video captioning [34], virtual reality [42], video compression[16], human-robot interaction [14, 26, 41], etc., owing to it's ability to prioritize the video information across space and time.

Initial efforts on the problem of video saliency predictions were limited to visual-only input. For instance, larger datasets like DHF1K [51] 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 [36] , [50] indicate the impact of audio in directing the human gaze. To understand the role of audio, comprehensive eye-tracking analysis [11, 28] 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 [44].

A series of audio-visual saliency prediction methods followed. Tavakoli *et al.* [47] 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.* [30] predicts audio saliency by cannonical correlation of visual and audio features and then fuses it with deep learning-based saliency models. STAViS [49] extends the SUSiNet's [21] visual saliency model and investigates three different ways to fuse the audio modality. Some recent efforts, have focused on face saliency, i.e. predicting the salient face in multi face videos. Liu *et al.* [22] concatenates features from three different streams (one for faces, one for visual embedding and one for audio) and decodes saliency maps using it. In [38] 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.

In this work, 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 (Figure 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 claims mentioned above. 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. Our 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.

2 RELATED WORKS

2.1 The role of audio in HVA

It is evident that audio matters in Human Visual Attention (HVA). Papai et al. [35] 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.* [6] 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 [9, 48] also verify the impact of audio signal on human attention. The work in [10, 24] 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.* [10] further suggest that in conversational video, increasing saliency of speakers' face greatly improves the model prediction. Other similar studies [11, 12, 43] 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

Initial deep learning based saliency prediction methods were limited to visual information. The methods can be largely classified in two categories. (a) The LSTM based models, which build on image-based saliency and aggregate frame-wise prediction using an LSTM [13, 52]. (b) 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 [18, 27]. Needless to say, the

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 [40] 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 dominated by the 3DConv based architectures. We rely primarily on the ViNet [18] model for our experiments, owing to its simplicity and decent performance.

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.* [47] used a 3D Resnet to encode both visual and audio information. They employ a simple concatenation operation on the encoded features. Chen *et al.* [5] also use concatenation operations, while using features from different visual hierarchies. STAViS [49] fuses the audio features onto the SUSiNet [21] visual encoder. They employ three different fusion methodologies namely cosine similarity, weighted inner product and bilinear transformation. Zhu *et al.* [55] 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.* [18] experiment with similar fusion methods to [21] on the ViNet backbone.

Some saliency prediction efforts have focused on conversational multi-face videos. Liu *et al.* [22, 38] employ multi-stream end-to-end trainable deep learning architectures. They propose a large scale MVVA dataset allowing efficient training. Several deep learning methods have also been explored [18, 22, 47].

Most of the aforementioned audio-visual saliency prediction methods claim that fusing audio leads to noticeable performance gains. Jain *et al.* [18] 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.

3 METHODOLOGY

We analyze the 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. 2).

For this study, we choose STAViS[49] and AViNet[18] networks that fuse spatio-temporal visual and auditory information to obtain a final saliency map.

3.1.1 STAViS : . We train STAViS [49] that extends the SusiNet [21] visual saliency model by fusing an audio modality. The visual branch consists of spatio-temporal module based on 3D-ResNet Blocks pre-trained on Kinetics-400 dataset [4]. This is followed by a Deeply Supervised Attention Module (DSAM), i.e., feature-wise multiplication of the output of each block and the attention map.

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 [49].

3.1.2 AViNet : We train AViNet, a U-Net like encoder-decoder network with a visual branch based on a S3D [53] backbone pre-trained on Kinetics-400 action recognition dataset [4]. 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 [18].

3.2 Audio Modules

To analyze the role of audio, we perform ablation across three different audio modules, i.e. SoundNet[1], VGG-Vox[7] and AVID[33]. These modules have shown significant performance in diverse correlated audio-visual tasks.

3.2.1 SoundNet : For sound representation, we employ SoundNet[1] 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[7] 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[7] dataset is fine-tuned for our task by end-to-end training of AVSP model.

3.2.3 AVID : . Furthermore to verify the role of audio in the aforementioned task, we employ AVID[33] 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

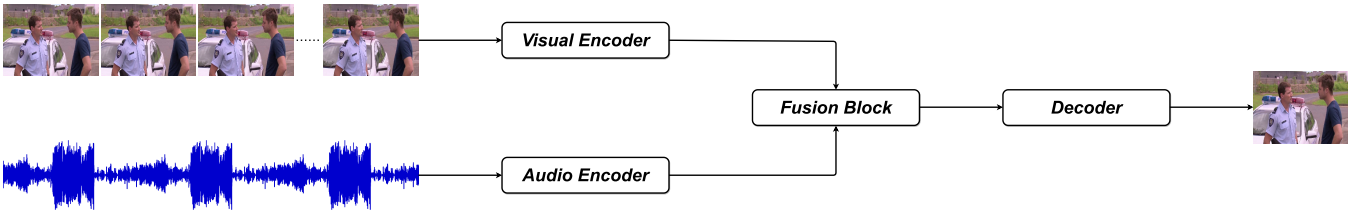

Figure 2: Audio-Visual Saliency Prediction model in general.

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 for our analysis, owing to their ability to generalize well across multiple domains, thereby leveraging multi-modal information.

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 [18, 49]. Bilinear fusion method captures the pairwise interactions across the feature dimensions.

We also performed our experiments with a simple concatenation technique used previously [18, 49]. 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-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 the works of Tao *et al.*[46], 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. 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.

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.* [37] 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. Each modality can be represented as a hypersphere with a mean feature norm as its radius. 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 hypersphere of a fixed radius. Planamente *et al.* [37] proposed a Relative Norm Alignment (RNA) loss that aims to align the mean feature norms of the two modalities. 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 resulting in faster convergence.

3.4 Regularization over Visual-Only Models

Dropout [45] is a regularization technique to ameliorate over-fitting in neural networks. Specifically, during the training phase, dropout randomly discards 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 4)

4 EXPERIMENT

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 DHF1K. [51] 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.

4.1.2 Hollywood-2. [25] 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.

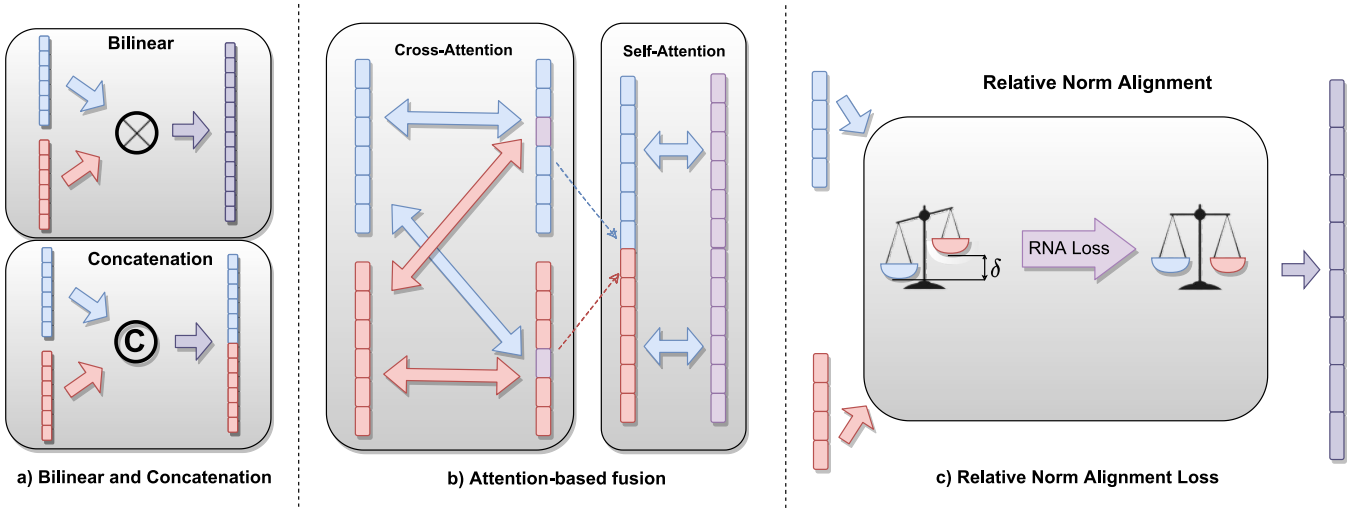

Figure 3: Different Fusion Techniques

4.1.3 UCF Sports. [39] 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.

4.1.4 DIEM. [31] 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.

4.1.5 Coutrot. [9, 23] 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.

4.1.6 SumMe. [15] 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.

4.1.7 AVAD. [29] 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.

4.1.8 ETMD. [20] dataset consists of 12 videos from 6 different Hollywood movies. The corresponding eye-tracking data are annotated by 10 observers.

4.1.9 MVVA. [22] 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.

4.2 Training procedure

For training AViNet, a similar training procedure is incorporated, as discussed by Jain *et al.*[18]. 32 consecutive frames are randomly selected from each clip of the dataset with their corresponding audio stream. Each frame is resized to 224X384 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 (KLDiv) 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 [49]. 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 [3]. 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(SIM), 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.

5 RESULTS AND DISCUSSIONS

Extensive experiments are performed to examine the role of audio in current AVSP models. The study is carried out on ten different audio-visual saliency datasets. We also attempt to investigate the cause for incremental gains in current AVSP models over the visual-only models.

5.1 Audio-visual Dataset

5.1.1 Role of Audio in SOTA models : To analyse 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 1, we find that the model inferred with different sound signals gives notably similar performance, thus showing an agnostic behaviour of both the SOTA models with audio on all the audio visual datasets. This behaviour suggest that the SOTA models are unable to utilize audio module at it's best and limits the performance of AVSP models. To this end, we tried different techniques to incorporate audio in a better way. We choose ViNet (being an outperforming model over STAViS) as base model for all our further experiments.

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[46], audio-visual objects learning[33], etc. Table **??** 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. To this end, we adopt different fusion techniques to integrate the audio and visual features in a better way.

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 2 compares the effect of different fusion techniques on network's performance. A minimalistic jitter in results suggests 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[37] to bring out norm balance and leverage audio in a better way. Table 3 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.

5.1.4 Regularization by Audio branch : 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 simple regularization technique on Visual model can help to learn the saliency of similar or higher precision. Table 4 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. 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.

5.2 Visual Only Datasets

The above hypothesis is validated by conducting the same set of experiments on visual-only datasets with audio input set to a random vector. Table 7 clearly shows that, AViNet shows a similar to better performance as compared to ViNet. Particularly in UCF dataset, a significant gain can be observed, leading to SOTA performance by just adding a random audio module. The dropout is further seen to inhibit this behaviour. Thus we bring into light that previously claimed audio-visual models don't incorporate audio but end up regularizing the visual module.

6 CONCLUSION

Our work presents a comprehensive analysis to underline the role of audio in current deep AVSP methods. 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. 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.

Our 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

Table 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

Table 2: 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 bilinear, 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

Table 3: Mean and standard deviation of feature norm before and after applying RNA Loss

	AViNet with BiLinear 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

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 [8] 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.

Overall, we believe the experiments presented in this paper will help the community reflect upon the role of audio in the current

Table 4: 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: 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

Table 6: Results of all the experiments discussed, on a recently proposed large-scale multi-face saliency dataset - MVVA[22]. A similar behaviour can be observed.

	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

Table 7: 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

research landscape, identify the shortcomings, and help build improved AVSP models.

REFERENCES

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

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

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

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

[5] Jiazhong Chen, Qingqing Li, Hefei Ling, Dakai Ren, and Ping Duan. 2021. Audio-visual saliency prediction via deep learning. *Neurocomputing* (2021).

[6] Yanxiang Chen, Tam V Nguyen, Mohan Kankanhalli, Jun Yuan, Shuicheng Yan, and Meng Wang. 2014. Audio matters in visual attention. *IEEE Transactions on Circuits and Systems for Video Technology* (2014).

[7] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. 2018. VoxCeleb2: Deep Speaker Recognition. In *Proc. Interspeech 2018.*

[8] Mert Cokelek, Nevrez Imamoglu, Cagri Ozcinar, Erkut Erdem, and Aykut Erdem. 2021. Leveraging Frequency Based Salient Spatial Sound Localization to Improve 360° Video Saliency Prediction. In *2021 17th International Conference on Machine Vision and Applications (MVA).*

[9] Antoine Coutrot and Nathalie Guyader. 2014. How saliency, faces, and sound influence gaze in dynamic social scenes. *Journal of vision* (2014).

[10] Antoine Coutrot and Nathalie Guyader. 2015. An efficient audiovisual saliency model to predict eye positions when looking at conversations. In *2015 23rd European Signal Processing Conference (EUSIPCO).*

[11] Antoine Coutrot, Nathalie Guyader, Gelu Ionescu, and Alice Caplier. 2012. Influence of soundtrack on eye movements during video exploration. *Journal of Eye Movement Research* (2012).

[12] Antoine Coutrot, Nathalie Guyader, Gelu Ionescu, and Alice Caplier. 2014. Video viewing: do auditory salient events capture visual attention? *annals of telecommunications-annales des télécommunications* (2014).

[13] Richard Droste, Jianbo Jiao, and J Alison Noble. 2020. Unified image and video saliency modeling. In *European Conference on Computer Vision.*

[14] Joao Filipe Ferreira and Jorge Dias. 2014. Attentional mechanisms for socially interactive robots–a survey. *IEEE Transactions on Autonomous Mental Development* (2014).

[15] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, and Luc Van Gool. 2014. Creating summaries from user videos. In *European conference on computer vision.*

[16] Hadi Hadizadeh and Ivan V Bajić. 2013. Saliency-aware video compression. *IEEE Transactions on Image Processing* (2013).

[17] Laurent Itti, Christof Koch, and Ernst Niebur. 1998. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence* (1998).

[18] Samyak Jain, Pradeep Yarlagadda, Shreyank Jyoti, Shyamgopal Karthik, Ramanathan Subramanian, and Vineet Gandhi. 2020. Vinet: Pushing the limits of visual modality for audio-visual saliency prediction. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS).*

[19] Pattaraporn Khuwuthyakorn, Antonio Robles-Kelly, and Jun Zhou. 2010. Object of interest detection by saliency learning. In *European conference on Computer vision.*

[20] Petros Koutras and Petros Maragos. 2015. A perceptually based spatio-temporal computational framework for visual saliency estimation. *Signal Processing: Image Communication* (2015).

[21] Petros Koutras and Petros Maragos. 2019. Susinet: See, understand and summarize it. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops.*

[22] Yufan Liu, Minglang Qiao, Mai Xu, Bing Li, Weiming Hu, and Ali Borji. 2020. Learning to Predict Salient Faces: A Novel Visual-Audio Saliency Model. In *European Conference on Computer Vision.*

[23] Matei Mancas, Vincent P Ferrera, Nicolas Riche, and John G Taylor. 2016. *From Human Attention to Computational Attention.*

[24] Pierre Marighetto, Antoine Coutrot, Nicolas Riche, Nathalie Guyader, Matei Mancas, Bernard Gosselin, and Robert Laganiere. 2017. Audio-visual attention: Eye-tracking dataset and analysis toolbox. In *2017 IEEE International Conference on Image Processing (ICIP).*

[25] Marcin Marszalek, Ivan Laptev, and Cordelia Schmid. 2009. Actions in context. In *2009 IEEE Conference on Computer Vision and Pattern Recognition.*

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

[27] Kyle Min and Jason J Corso. 2019. Tased-net: Temporally-aggregating spatial encoder-decoder network for video saliency detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision.*

[28] Xiongkuo Min, Guangtao Zhai, Zhongpai Gao, Chunjia Hu, and Xiaokang Yang. 2014. Sound influences visual attention discriminately in videos. In *2014 Sixth International Workshop on Quality of Multimedia Experience (QoMEX).*

[29] Xiongkuo Min, Guangtao Zhai, Ke Gu, and Xiaokang Yang. 2016. Fixation prediction through multimodal analysis. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* (2016).

[30] Xiongkuo Min, Guangtao Zhai, Jiantao Zhou, Xiao-Ping Zhang, Xiaokang Yang, and Xinping Guan. 2020. A multimodal saliency model for videos with high audio-visual correspondence. *IEEE Transactions on Image Processing* (2020).

[31] Parag K Mital, Tim J Smith, Robin L Hill, and John M Henderson. 2011. Clustering of gaze during dynamic scene viewing is predicted by motion. *Cognitive computation* (2011).

[32] KL Bhanu Moorthy, Moneish Kumar, Ramanathan Subramanian, and Vineet Gandhi. 2020. Gazed–gaze-guided cinematic editing of wide-angle monocular video recordings. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems.*

[33] Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. 2021. Audio-Visual Instance Discrimination with Cross-Modal Agreement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).*

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

[35] Márta Szabina Pápai and Salvador Soto-Faraco. 2017. Sounds can boost the awareness of visual events through attention without cross-modal integration. *Scientific reports* (2017).

[36] David R Perrott, Kourosh Saberi, Kathleen Brown, and Thomas Z Strybel. 1990. Auditory psychomotor coordination and visual search performance. *Perception & psychophysics* (1990).

[37] Mirco Planamente, Chiara Plizzari, Emanuele Alberti, and Barbara Caputo. 2022. 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.*

[38] Minglang Qiao, Yufan Liu, Mai Xu, Xin Deng, Bing Li, Weiming Hu, and Ali Borji. 2021. Joint Learning of Visual-Audio Saliency Prediction and Sound Source Localization on Multi-face Videos. *arXiv preprint arXiv:2111.08567* (2021).

[39] Mikel D Rodriguez, Javed Ahmed, and Mubarak Shah. 2008. Action mach a spatio-temporal maximum average correlation height filter for action recognition. In *2008 IEEE conference on computer vision and pattern recognition.*

[40] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention.*

[41] Guido Schillaci, Saša Bodiroža, and Verena Vanessa Hafner. 2013. Evaluating the effect of saliency detection and attention manipulation in human-robot interaction. *International Journal of Social Robotics* (2013).

[42] Vincent Sitzmann, Ana Serrano, Amy Pavel, Maneesh Agrawala, Diego Gutierrez, Belen Masia, and Gordon Wetzstein. 2018. Saliency in VR: How do people explore virtual environments? *IEEE transactions on visualization and computer graphics* (2018).

[43] Guanghan Song, Denis Pellerin, and Lionel Granjon. 2011. Sound effect on visual gaze when looking at videos. In *2011 19th European Signal Processing Conference.*

[44] Guanghan Song, Denis Pellerin, and Lionel Granjon. 2013. Different types of sounds influence gaze differently in videos. *Journal of Eye Movement Research* (2013).

[45] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research* (2014).

[46] Ruijie Tao, Zexu Pan, Rohan Kumar Das, Xinyuan Qian, Mike Zheng Shou, and Haizhou Li. 2021. Is someone speaking? exploring long-term temporal features for audio-visual active speaker detection. In *Proceedings of the 29th ACM International Conference on Multimedia.*

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

[48] Antigoni Tsiami, Athanasias Katsamanis, Petros Maragos, and Argiro Vatakis. 2016. Towards a behaviorally-validated computational audiovisual saliency model. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).*

[49] Antigoni Tsiami, Petros Koutras, and Petros Maragos. 2020. STAViS: Spatio-Temporal AudioVisual Saliency Network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).*

[50] Jean Vroomen and Beatrice de Gelder. 2000. Sound enhances visual perception: cross-modal effects of auditory organization on vision. *Journal of experimental psychology: Human perception and performance* (2000).

[51] Wenguan Wang, Jianbing Shen, Fang Guo, Ming-Ming Cheng, and Ali Borji. 2018. Revisiting video saliency: A large-scale benchmark and a new model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*

[52] Xinyi Wu, Zhenyao Wu, Jinglin Zhang, Lili Ju, and Song Wang. 2020. Salsac: A video saliency prediction model with shuffled attentions and correlation-based convlstm. In *Proceedings of the AAAI Conference on Artificial Intelligence.*

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

[54] Tong Yubing, Faouzi Alaya Cheikh, Fahad Fazal Elahi Guraya, Hubert Konik, and Alain Trémeau. 2011. A spatiotemporal saliency model for video surveillance. *Cognitive Computation* (2011).

[55] Dandan Zhu, Defang Zhao, Xiongkuo Min, Tian Han, Qiangqiang Zhou, Shaobo Yu, Yongqing Chen, Guangtao Zhai, and Xiaokang Yang. 2021. Lavs: A Lightweight Audio-Visual Saliency Prediction Model. In *2021 IEEE International Conference on Multimedia and Expo (ICME).*