ViNet: Pushing the limits of Visual Modality for Audio-Visual Saliency Prediction

Samyak Jain[1], Pradeep Yarlagadda[1], Shreyank Jyoti[1], Shyamgopal Karthik[1],
Ramanathan Subramanian[2], Vineet Gandhi[1]

Abstract— **We propose the ViNet architecture for audio-visual saliency prediction. ViNet is a fully convolutional encoder-decoder architecture. The encoder uses visual features from a network trained for action recognition, and the decoder infers a saliency map via trilinear interpolation and 3D convolutions, combining features from multiple hierarchies. The overall architecture of ViNet is conceptually simple; it is causal and runs in real-time (60 fps). ViNet does not use audio as input and still outperforms the state-of-the-art audio-visual saliency prediction models on nine different datasets (three visual-only and six audio-visual datasets). ViNet also surpasses human performance on the CC, SIM and AUC metrics for the *AVE* dataset, and to our knowledge, it is the first network to do so. We also explore a variation of ViNet architecture by augmenting audio features into the decoder. To our surprise, upon sufficient training, the network becomes agnostic to the input audio and provides the same output irrespective of the input. Interestingly, we also observe similar behaviour in the previous state-of-the-art models [1] for audio-visual saliency prediction. Our findings contrast with previous works on deep learning-based audio-visual saliency prediction, suggesting a clear avenue for future explorations incorporating audio in a more effective manner. The code and pre-trained models are available at `https://github.com/samyak0210/ViNet`.**

I. INTRODUCTION

Video saliency prediction focuses on understanding and modeling human visual attention (HVA) while viewing a dynamic scene (determining where and what people pay attention to given visual stimuli). HVA empowers primates to analyze/interpret the complex surroundings rapidly, and naturally, we would like to extend these abilities to machines/robots. For instance, a robot that orients its eyes like humans gives impressions of an intelligent behaviour [2]. Moreover, it may allow the robot to orient towards regions of the visual scene that are likely to be relevant. Upon compiling the *ground truth* regarding where viewers gaze in the scene via eye-tracking hardware, saliency prediction (SP) aims to mimic HVA given a novel video computationally. Previous works have shown that SP is valuable in a variety of applications like human-robot interaction [3], stream compression [4], video captioning [5], automated cinematic editing [6], *etc*.

Video SP models primarily employ visual information to predict gaze. Larger datasets like *DHF1K* [7] discard audio during ground truth collection, and ask users to look at *silent* videos. End-to-end deep saliency models are then trained using only visual information. State-of-the-art video SP models largely depend on Long Short-Term Memory (LSTM) networks to encode temporal dependencies [8], [9], [10]. These models build on image-based saliency and aggregate frame-wise prediction using an LSTM. Since both spatial decoding and temporal aggregation are performed separately, LSTM models cannot collectively leverage Spatio-temporal information, shown to be beneficial for video SP [11].

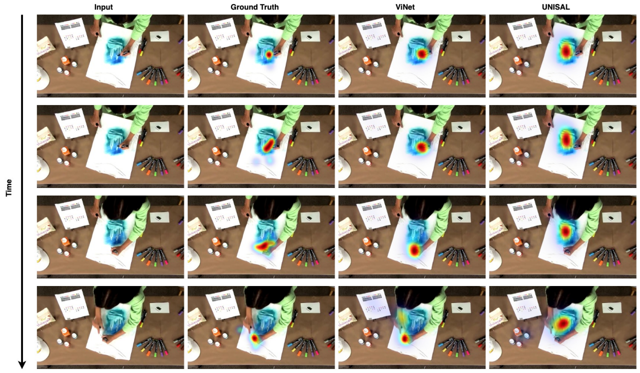

Fig. 1. The core of our approach is a strong visual-only model ViNet. Here, we compare ViNet (third column) with state-of-the-art UNISAL model [8] (fourth column). Note that ViNet better captures the action, while UNISAL focuses on objectness. In this example, ViNet focuses on the region being drawn, whereas UNISAL focuses on the completed portion. Best viewed in color and under zoom.

To this end, we propose a novel fully convolutional encoder-decoder architecture called ViNet for visual saliency detection. ViNet takes a set of frames as input and predicts a saliency map for the last frame. Following the methodology adopted in [11], it then employs a sliding window approach to predict saliency for the entire video. ViNet takes features learned from an action recognition network from multiple hierarchies, fuses them in a UNet [12] like fashion, and outputs a saliency map using trilinear interpolations and 3D convolutions. The strength of ViNet is that it only comprises commonly used components, resulting in a minimal and conceptually simple model which is easy to train and interpret. ViNet is causal, runs in real-time, and surpasses the state-of-the-art on three popular vision-only saliency prediction datasets (a motivating example is illustrated in Fig. 1). At the time of submission, ViNet is also the top-ranked model on the private test-set of *DHF1K*, the most diverse video saliency prediction benchmark. Interestingly, ViNet also achieves state-of-the-art results on six audio-visual saliency datasets without using any audio information.

[1] CVIT, KCIS, International Institute for Information Technology, Hyderabad `samyak.j@research.iiit.ac.in`
[2] Indian Institute of Technology, Ropar

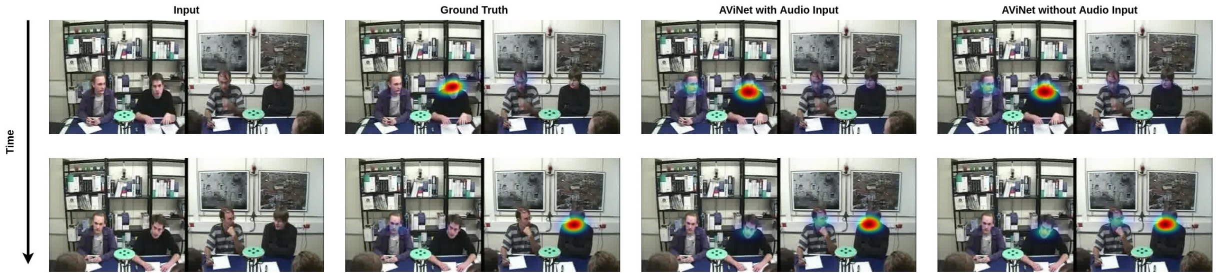

Fig. 2. Sample frames from *Coutrot-2* database with the corresponding ground-truth. The predicted saliency maps of AViNet with and without passing audio input turn out to be the same.

More fundamentally, discarding audio information contrasts with our real-life behaviour, where we simultaneously perceive visual and audio modalities. Cognitive studies confirm that auditory and visual cues are correlated and jointly contribute to human attention [13]. Coutrot *et al.* [14] collect the human gaze on the same set of videos with and without the original soundtrack and observe that the soundtrack significantly affects the attention models in human perception, even when using a monophonic stimuli [14]. Consequently, recent efforts have explored multi-modal (audio-visual) video SP [15], [1], and claim audio as a strong cue for SP.

Consequently, we experiment with an audio-visual saliency prediction model obtained by augmenting ViNet with an audio branch. The resulting architecture called AViNet is end-to-end trainable and uses pre-trained audio features from SoundNet [16]. We explore two fusion strategies, similar to [1], [15] *i.e.* simple concatenation and bilinear fusion. We observe that when compared to ViNet, AViNet gives nil or marginal improvements on most audio-visual SP datasets. Our results suggest that current audio-visual saliency models [15], [1] are not optimal on the visual modality. Furthermore, when we dig deeper, we find out that the audio-visual network learns to ignore the audio signal entirely and gives the same result even while sending a zero vector as audio or by sending an unrelated random audio file (Fig. 2). Surprisingly, we observe the same behaviour with STAViS [1], the current state-of-the-art audio-visual saliency prediction model. Our findings contrast to the prevalent claims that the audio acts as a strong cue for SP. Overall, we make the following research contributions:

- We propose a novel visual-only architecture called ViNet for video saliency detection. Our model uses commonly known deep learning components/ideas, and the contributions are in their efficient amalgamation. We back the proposed architecture with thorough ablation studies.
- We present a comprehensive analysis on ten different datasets (three visual and seven audio-visual datasets). Our model achieves solid performance gains over the current state-of-the-art.
- We carefully explore the role of audio and find that the visual-only model almost recovers the underlying performance. Furthermore, the strategies mentioned in existing literature end up learning a prediction model agnostic to audio. This motivates the need for exploring novel architectures for audio-visual fusion for SP and possibly carefully curating datasets where audio plays a significant role.

II. Related Works

A. Video Saliency

The recent landscape in video saliency prediction is dominated by the end-to-end trainable deep networks. The availability of large datasets like *Hollywood-2* [17] and *DHF1K* [7] have been instrumental in this progress. *Hollywood-2* is the largest dataset, however, its content is limited to human actions and movie scenes. *DHF1K* is considered the most diverse and challenging dataset for saliency detection.

Majority of the recent approaches rely on an LSTM based architecture for sequential fixation prediction over successive frames. Wang *et al.* [7] combine frame-level image features using a ConvLSTM. SalEMA [10] model recurrence using a temporal exponential moving average (EMA) operation over the convolutional layer. They show such a simple moving average-based approach matches the performance achieved using a ConvLSTM. SALSAC [9] adds further complexity to basic ConvLSTM architecture through a shuffled multi-level attention module and a frame correlation module. STRA-Net [18] learns an alignment module, and then aligned frames are sent into a Bi-ConvLSTM. [19] propose a novel construction of LSTM (2C-LSTM) with two sub-networks to focus on objectness and motion, respectively. UNISAL [8] is a unified image and video saliency prediction model that uses MobileNet to extract spatial features and LSTMs for encoding temporal information. The method heavily relies on domain adaptive prior maps (different prior maps for image and video domains), domain adaptive batch-normalization, etc. Several of these video saliency prediction architectures [8], [7] borrow and extend ideas (hierarchical features, transfer learning, multi-branch architectures, etc.) from the models trained for static image saliency prediction [20], [21].

3D convolutional architectures have also been explored for the task. These methods typically rely on action detection networks as their backbone. TASED-Net [11] uses S3D as an encoder to extract spatial features while jointly aggregating all the temporal information in order to produce a single full-resolution prediction map. They use transposed

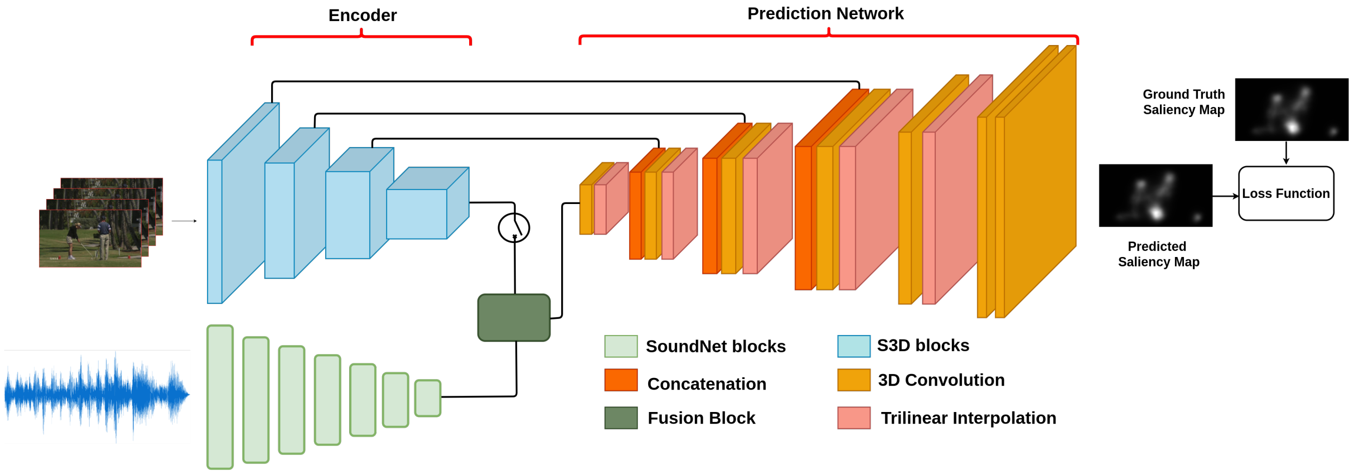

Fig. 3. AViNet Architecture overview. Removing the audio branch, the resulting architecture is ViNet.

convolution layers with auxiliary pooling (a variation of max-unpooling layers) for spatial upscaling in the decoder. Bellitto *et al.* [22] use multiple decoders for features encoded at different levels to obtain multiple saliency instances that are finally combined to obtain final output saliency maps. [22] is inspired by the DVA image saliency model [23]. A combination of 3D convolutions and recurrent architecture has also been explored [24]. STSConvNet [25] explicitly computes optical flow and fuses the optical with the visual features into two-stream convolutional architecture.

In contrast, our ViNet method is a straightforward encoder-decoder architecture exploiting basic ideas of spatial hierarchy, feature concatenation, skip connections, trilinear upsampling, and 3D convolutions. It uses pretrained features from a network trained for action recognition as a backbone and is void of any explicit inputs like optical flow or any extra modules for detecting objectness, motion, attention, etc.

B. Audio-Video Saliency

Research in cognitive neuroscience has led to interesting findings about audiovisual integration. If you ever watched a ventriloquist in action, you would agree how they trick our visual stimuli to guide the perceived location of the sound (and where we look at). Ventriloquist turns to face the puppet, they attend the puppet, use a different voice for the puppet, and make it seem that it is the puppet that is talking (although the sound is being generated from their stomach). McGurk effect [26], pip and pop effect [27], unity assumptions [28] are other examples of how we jointly integrate and perceive visual and audio modalities. Coutrot *et al.* [14], [29], [30] present some interesting studies on the influence of soundtrack on eye movements during video exploration.

Application-specific attempts have been made for visual saliency and audio localization [31], [32], [33]. The fusion of handcrafted attention models and pre-trained deep image-level models using canonical correlation analysis has been explored [34], [35]. However, only a couple of attempts have been made towards an end-to-end deep learning-based audio-visual saliency fixation prediction. Tavakoli et al. [15] trains two independent networks for the two modalities (audio and visual data), and their outputs are simply concatenated as a late fusion scheme. They use 3DResNet as the backbone for both modalities. STAViS [1] extends the SUSiNet [36] visual saliency model and investigates three different ways to fuse the audio modality.

Significant efforts have been made in the direction of self-supervised learning and representation learning exploiting audio-visual data. SoundNet [16] leverage the natural synchronization between vision and sound to learn an acoustic representation. They use a student-teacher training procedure to transfer discriminative visual knowledge (large-scale visual recognition) into the sound modality. On similar lines, audiovisual correspondence has been exploited for the task of cross-modal retrieval [37], sound classification [16], [38], sound localization in images [37], [39], scene analysis [40], temporal event localization [41] etc.

III. Proposed Architecture

We propose an end-to-end architecture visual-only model called ViNet. It is a fully 3D-convolutional encoder-decoder architecture that predicts the saliency for the last frame of the corresponding set of sequential frames. Then we present an audio-visual saliency detection model called AViNet that fuses the visual features from ViNet and audio features from SoundNet. Fig. 3 displays an overview of the architecture.

A. Backbone

The architecture uses the S3D network [42] as the video encoder. We use the model pre-trained on the Kinetics dataset which is an action-recognition dataset. We use S3D since it consists of 3D convolutional layers which efficiently encodes the spatio-temporal information. Moreover, it is light-weight and pre-trained on a large dataset, making it fast and effective for transfer-learning. It consists of 4 convolutional blocks base1, base2, base3 and base4 that provides outputs X_1,X_2,X_3 and X_4 in different spatial and temporal scales.

Clip Size (T)	CC	SIM	NSS
8	0.4978	0.363	2.8221
16	0.5112	0.378	2.9067
32	0.5212	**0.3881**	**2.9565**
48	**0.5231**	0.3807	2.9477

TABLE I
VALIDATION RESULTS ON VARYING CLIP SIZE FOR TRAINING VINET ON EMPHDHF1K.

Model Architecture	CC	SIM	NSS
Without Hierarchy	0.5002	0.361	2.7371
With Hierarchy	0.5212	0.3881	2.9565

TABLE II
VALIDATION RESULTS OF VINET WITH AND WITHOUT HIERARCHY ON *DHF1K*.

X_1,X_2 and X_3 are referred as multi-level features that are extracted at three-levels of hierarchy. The input to the encoder is a video clip $x_{clip} \in R^{3\times T_0\times H_0\times W_0}$, where T_0 is 32. It generates a lower-resolution activation map $X_4 \in R^{C\times T\times H\times W}$, where $C = 1024$ and $T, H, W = \frac{T_0}{8}, \frac{H_0}{32}, \frac{W_0}{32}$.

For audio representation, we employ SoundNet [16], which is trained for audio/sound based scene classification. We pre-process the audio data similar to the STAViS [1] (section 3.2). The sound module takes 1D pre-processed audio feature as input, $y_{audio} \in R^{1\times\hat{T}\times 1}$ and outputs audio features $A \in R^{1024\times 3\times 1}$.

B. Audio-Visual Fusion

Inspired by the recent works on audio-visual saliency prediction [1], [15], we explore two types of fusion techniques. First is a simple concatenation of encoded audio and video features which was used in [15]. We repeat the audio features to match the dimensions of visual features and combine them across the channel dimension. Then we apply 1 × 1 Convolution to reduce the number of channels.

Secondly, we applied bilinear fusion which has been used in [1]. The visual features are first passed through Max Pool to reduce the spatial and temporal dimension and then collapsed to represent it as a vector $x_1 \in R^{1024\times x_0}$. Similarly, the audio features are collapsed as a vector $x_2 \in R^{1024\times y_0}$. The bilinear fusion is defined as

$$y = x_1^T A x_2 + b \tag{1}$$

where $A \in R^{x_0\times x\times y_0}$ and $b \in R^{x\times 1}$ are parameters and x is the desired output dimension.

C. Prediction Network

The Prediction Network consists of 6 decoding layers consisting of 3D convolutional and upsampling layers. For ViNet, the input to the Prediction Network is the X_4 features from the Backbone and X_3,X_2, and X_1 are passed in using skip connections, respectively. In the case of AViNet, the audio features are fused with X_4 and then sent to the decoder (skip connections are made similarly).

D. Evaluation

Both ViNet and AViNet follow a sliding window approach to generate a saliency map for all frames in the video. Given a window size of T frames, we predict saliency map S_t at time step t by taking $F_{t-T+1}, ...F_t$ sequence of frames as input. To enable prediction in the first T frames, we simply repeat the first frame of the video at the start. A single inference of ViNet takes around 0.016 seconds (62.5 fps) to generate a saliency map, with T = 32 frames.

IV. EXPERIMENTS AND RESULTS

A. Datasets

1) Visual Datasets: The three most popular visual-only datasets in video saliency are *DHF1K*, *Hollywood-2*, and *UCF-Sports* [45]. We carry out the tests and comparisons on these three datasets.

DHF1K [7] contains 1000 videos where 700 videos are for training and 100 for validation. A test set of 300 videos is also released, however, without public ground truth. All our experiments and analysis are based on this dataset since it is the most general and diverse dataset.

Hollywood-2 [17] is the largest video saliency prediction dataset in terms of the number of videos, consisting of 1707 videos. The dataset is focused on human actions. The videos in this dataset are short video sequences from a set of 69 Hollywood movies, containing 12 different human action classes, ranging from answering the phone, eating, driving, running and etc. We use the standard split of 823 training videos and 884 test videos.

UCF-Sports [45] dataset consists of 150 videos focusing on human actions in sports. We use a standard split with 103 videos for training and 47 videos for testing.

2) Audio-Visual Datasets: There are seven audio-visual datasets in video saliency: *DIEM*, *Coutrot1*, *Coutrot2*, *AVAD*, *ETMD*, *SumMe*, and *AVE* dataset. We carry out the tests and comparisons on all these seven datasets.

DIEM [46] consists of 81 movie clips of varying genres. They sourced from publicly accessible repositories, including advertisements, documentaries, game trailers, movie trailers, music videos, news clips, and time-lapse footage. It consists of 64 training videos and 17 test videos.

Coutrot databases [29], [30] are split into *Coutrot1* and *Coutrot2*. *Coutrot1* contains 60 clips with dynamic natural scenes split into four visual categories: one/several moving objects, landscapes, and faces. *Coutrot2* contains 15 clips of 4 persons in a meeting and the corresponding eye-tracking data from 40 persons.

AVAD dataset [34] contains 45 short clips of 5-10 sec duration with several audio-visual scenes, e.g., dancing, guitar playing, birds singing, etc.

ETMD dataset [47] contains 12 videos from six different hollywood movies.

SumMe dataset [48] 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.

	DHF1K					*Hollywood-2*					*UCF-Sports*				
	CC	sAUC	AUC	NSS	SIM	CC	sAUC	AUC	NSS	SIM	CC	sAUC	AUC	NSS	SIM
SALEMA [10]	0.449	0.667	0.890	2.57	0.466	0.613	0.708	0.919	3.18	0.487	0.544	0.740	0.906	2.63	0.431
ACLNet [43]	0.434	0.601	0.890	2.35	0.315	0.623	0.757	0.913	3.08	0.542	0.510	0.744	0.897	2.56	0.406
STRA-Net [44]	0.458	0.663	0.895	2.55	0.355	0.662	0.774	0.923	3.47	0.536	0.593	0.751	0.910	3.01	0.479
SALSAC [9]	0.479	0.697	0.896	2.67	0.357	0.670	0.712	0.931	3.35	0.529	0.671	0.806	0.926	3.52	0.534
TASED-Net [11]	0.470	0.712	0.895	2.66	0.361	0.64	0.768	0.918	3.30	0.507	0.582	0.752	0.899	2.92	0.469
UNISAL [8]	0.490	0.691	0.901	2.77	0.390	0.673	0.795	0.934	3.90	0.542	0.644	0.775	0.918	3.38	0.523
ViNet	0.510	0.728	0.908	2.87	0.381	0.693	0.813	0.930	3.73	0.550	0.673	0.810	0.924	3.62	0.522

TABLE III

COMPARISON RESULTS ON THE *DHF1K*, *Hollywood-2* AND *UCF-Sports* TEST SETS. THE BEST SCORES ARE SHOWN IN RED AND SECOND BEST SCORES IN BLUE.

AVE dataset [15] consists of 150 hand-picked video sequences from *DIEM*, *Coutrot1* and *Coutrot2* datasets. The videos are divided into three categories - Nature, Social Events, and Miscellaneous. The dataset consists of 92 training videos, 29 validation, and 29 test sequences.

B. Experimental Setup

For training ViNet, clips with T consecutive frames were randomly selected from the dataset. Each frame is resized to 224×384 and trained with a batch size of 8. The optimizer used is Adam, and the learning rate is set to be 1e-4. The network is initially trained on the *DHF1K* dataset. The validation set of *DHF1K* is used for early stopping. The trained model is then fine-tuned for *Hollywood-2* and *UCF-Sports* dataset using their respective training sets. The test sets of *Hollywood-2* and *UCF-Sports* are used for early stopping.

For our audio-visual extension AViNet, weights of ViNet pre-trained on *DHF1K* are used and fine-tuned on the audio-visual datasets. For *DIEM*, the standard split provided in the literature is used. For other datasets, there are no standard splits defined, so we evaluated our model on three different splits defined by [1] and report the average metric values across various splits. For evaluating on *AVE* dataset, we fine-tune the model using its training set and use its validation for early stopping.

C. Loss Function

We use the Kullback-Leibler divergence as the loss function, which is often used in saliency prediction tasks. KLDiv is an information-theoretic measure of the difference between two probability distributions:

$$KLdiv(P,Q) = \sum_i Q_i \log(\epsilon + \frac{Q_i}{P_i + \epsilon}), \tag{2}$$

where P, Q are predicted and ground truth maps respectively and ϵ is a regularization term.

D. Ablation studies

We present ablations studies that motivated our design choices in the ViNet model. All the ablations in this section are performed with training on the *DHF1K* training set and evaluation on its validation set. We examine the effects of (a) changing the clip size, (b) using multi-level features, (c) replacing upsampling with transpose convolutions, and (d) applying different concatenation techniques for fusing hierarchical features. Table I illustrates the results on varying the clip size of the input and using clips of size 32 frames gave the best results. Ablation results by using hierarchical features can be found in Table II. It clearly indicates that using multi-level features adds up to the performance. We also use transpose convolution instead of trilinear upsampling to increase the spatial dimension, but CC decreased to 0.5178 from 0.5212. The multi-level features extracted from the backbone are concatenated at each decoder block. We tried two ways of concatenating features - across temporal dimension and channel dimension. We observed that they gave a similar performance; therefore, we went ahead with the former approach due to fewer trainable parameters.

E. Comparison with state-of-the-art

a) Visual-Only datasets: We quantitatively compare our model with the top six state-of-the-art models on *DHF1K*, *Hollywood-2*, and *UCF-Sports* test set. Table III shows the results on all three datasets in terms of CC, sAUC, AUC, NSS, and SIM metrics. We can observe that ViNet outperforms all the state-of-the-art models on the *DHF1K* dataset. ViNet also achieves top results on most metrics on *Hollywood-2* and *UCF-Sports* datasets. At the time of the submission, ViNet is the top-performing model on the *DHF1K* challenge (evaluated on the private test set)[1]. We show a qualitiative example in Fig. 4 where we see that ViNet is able to produce much more accurate saliency maps as compared to TASED-Net and STAViS.

b) Audio-Visual Datasets: The comparison of ViNet and AViNet models with state-of-the-art methods on audio-visual datasets are presented in Table IV and V. We also present results on ViNet(NF) baseline model, which is a trained of *DHF1K* dataset and not fine-tuned further on audio-visual datasets. The ViNet, ViNet(NF), ACLNet and TASED-Net models are trained without using any audio information. STAViS and AViNet models make use of the audio modality, both during training and inference. AViNet(B) and AViNet(C) present the two fusion methodologies discussed above *i.e.* concatenation and bilinear fusion respectively.

[1] The challenge website can be found here https://mmcheng.net/videosal/

	DIEM					*Coutrot1*					*Coutrot2*				
	CC	sAUC	AUC	NSS	SIM	CC	sAUC	AUC	NSS	SIM	CC	sAUC	AUC	NSS	SIM
ACLNet [43]	0.522	0.622	0.869	2.02	0.427	0.425	0.542	0.85	1.92	0.361	0.448	0.594	0.926	3.16	0.322
TASED-Net [11]	0.557	0.657	0.881	2.16	0.461	0.479	0.58	0.867	2.18	0.388	0.437	0.611	0.921	3.17	0.314
STAViS [1]	0.579	0.674	0.883	2.26	0.482	0.472	0.584	0.868	2.11	0.393	0.734	0.71	**0.958**	5.28	**0.511**
ViNet(NF)	0.571	0.695	0.886	2.28	0.468	0.509	0.619	0.875	2.46	0.406	0.645	0.72	0.949	5.11	0.419
ViNet	0.626	**0.723**	0.898	2.47	0.483	0.551	0.633	0.886	2.68	0.423	0.724	0.739	0.95	5.61	0.466
AViNet(B)	**0.632**	0.719	**0.899**	**2.53**	**0.498**	**0.56**	0.635	**0.889**	**2.73**	0.425	**0.754**	**0.742**	0.951	**5.95**	0.493
AViNet(C)	0.631	0.720	0.897	2.50	0.497	0.556	**0.636**	0.887	2.68	**0.426**	0.753	0.743	0.951	5.81	0.486

TABLE IV

COMPARISON RESULTS ON THE *DIEM*, *Coutrot1* AND *Coutrot2* TEST SETS.

	AVAD					*ETMD*					*SumMe*				
	CC	sAUC	AUC	NSS	SIM	CC	sAUC	AUC	NSS	SIM	CC	sAUC	AUC	NSS	SIM
ACL-Net [43]	0.58	0.56	0.905	3.17	0.446	0.477	0.675	0.915	2.36	0.329	0.379	0.609	0.868	1.79	0.296
TASED-Net [11]	0.601	0.589	0.914	3.16	0.439	0.509	0.711	0.916	2.63	0.366	0.428	0.657	0.884	2.1	0.333
STAViS [1]	0.608	0.593	0.919	3.18	0.457	0.569	0.731	**0.931**	2.94	**0.425**	0.422	0.656	0.888	2.04	0.337
ViNet (NF)	0.665	0.651	0.923	3.67	0.501	0.544	0.719	0.924	2.92	0.404	0.455	0.687	0.893	2.35	**0.349**
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
AViNet(B)	0.674	0.658	0.927	3.77	0.491	**0.571**	0.733	0.928	**3.08**	0.406	0.463	0.692	0.897	2.41	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

TABLE V

COMPARISON RESULTS ON THE *AVAD*, *ETMD* AND *SumMe* TEST SETS.

The ViNet model significantly outperforms STAViS on most datasets across most metrics. Surprisingly, the ViNet(NF) model is already competitive, indicating that models trained on *DHF1K* can generalize well to other datasets. Moreover, the results clearly suggest that the visual-only modality, when exploited well, is able to recover most of the underlying performance on the current datasets, as compared to existing state-of-the-art models. The improvements obtained by AViNet(B) and AViNet(C) models over ViNet are marginal at best (with an exception on *Coutrot2* dataset, which is captured in highly specific settings). Hence, in contrast to previous works [1], [15], our experimental results do not indicate any clear benefit of incorporating audio in the prediction pipeline.

We also evaluate the performance of our models on the *AVE* dataset [15]. Although the *AVE* dataset is formed using the sequences from *DIEM*, *Coutrot1*, and *Coutrot2* datasets, it is an interesting dataset because (a) it provides a human upper bound and a lower bound using dataset biases and (b) it provides video level categorization. The upper bound is named Human Infinite (HI) and is computed by splitting the eye-movements of observers into two groups and assessing one group against the other (human vs. human performance). The lower bound is called the Mean Eye Position map (MEP) and is computed from the training sequences. It depicts the center-bias that a model may learn by training on the dataset. It is, hence, a robust lower-bound baseline.

ViNet model outperforms the state-of-the-art approaches on the *AVE* dataset by a significant margin, resonating with the observations on other datasets. Notably, ViNet is able to cross the HI upper bound on AUC-J, CC, and SIM metrics. We further provide a category-wise analysis of both our models on this dataset. It is evident from Table VII ViNet and AViNet give fairly similar performance across all three categories, giving solid gains over other methods.

Method	CC	SIM
AViNet(B)	0.9977	0.9979
AViNet(C)	0.9978	0.9990
STAViS	0.9980	0.9981

TABLE VI

CC AND SIM METRICS FOR THE AUDIO-VISUAL SALIENCY PREDICTIONS FOR AVINET(B), AVINET(C) AND STAVIS WITH AND WITHOUT AUDIO(SENDING ZEROS FOR AUDIO) ON *Coutrot2* DATASET. THE PREDICTIONS ARE NEARLY IDENTICAL AS REFLECTED IN THE METRICS.

F. The Impact of Audio

We conduct a simple experiment to investigate the role of audio in AViNet and STAViS models. We compare the output predictions obtained with original audio and by sending zeroed-out vector as audio (indicating the absence of audio). To our surprise, the network's prediction maps obtained with and without audio are nearly identical (as presented in Table VI). A qualitative example is shown in Fig. 2. This indicates that the network learns to be agnostic to audio and gives the same output irrespective of the audio input (zero vector, corresponding audio, or random audio). In summary, the current state-of-the-art audio-saliency models end up learning a visual-only model and that also explains the marginal differences with ViNet and AViNet models in our results (Table IV and Table V). Such marginal differences might arise due to different instances of training or possibly due to a slight variation in the number of parameters. A deeper exploration is left for future work. Finally, the

Cat.	Model Name	CC	sAUC	AUC	NSS	SIM
Nature	HI	0.669	**0.762**	0.866	**3.32**	**0.538**
	AViNet	0.649	0.729	0.895	2.37	0.515
	ViNet	**0.680**	0.735	**0.900**	2.47	**0.538**
	DAVE [15]	0.539	0.723	0.877	2.27	0.450
	ACLNet* [43]	0.517	0.723	0.884	2.03	0.401
	MEP	0.471	0.686	0.869	1.76	0.368
Soc Ev.	HI	0.655	0.759	0.855	**3.63**	0.516
	AViNet	**0.688**	**0.765**	**0.914**	2.96	0.536
	ViNet	**0.688**	0.760	0.910	2.88	**0.544**
	DAVE [15]	0.545	0.726	0.885	2.65	0.442
	ACLNet* [43]	0.449	0.683	0.869	2.02	0.359
	MEP	0.314	0.633	0.819	1.35	0.274
Misc.	HI	0.597	**0.748**	0.837	**3.23**	0.481
	AViNet	0.635	0.730	**0.898**	2.42	0.506
	ViNet	**0.636**	0.726	0.896	2.40	**0.509**
	Dave [15]	0.549	0.736	0.881	2.39	0.454
	ACLNet* [43]	0.456	0.683	0.852	1.84	0.378
	MEP	0.438	0.675	0.845	1.73	0.342
Overall	HI	0.644	**0.757**	0.854	**3.41**	0.514
	AViNet	0.655	0.744	0.901	2.55	0.516
	ViNet	**0.671**	0.742	**0.903**	2.60	**0.533**
	DAVE [15]	0.545	0.726	0.881	2.45	0.449
	ACLNet* [43]	0.475	0.700	0.870	1.98	0.379
	MEP	0.403	0.662	0.844	1.59	0.326

TABLE VII

PERFORMANCE OF VARIOUS MODELS ON AVE TEST SET CATEGORIES. HI REPRESENTS HUMAN INFINITE (HI) REPRESENTS UPPER PERFORMANCE BOUND AND MEAN EYE POSITION (MEP) REPRESENTS LOWER PERFORMANCE BOUND.

observations contrast with cognitive studies, which suggest clear differences in human gaze patterns when the videos are watched with or without audio [14]. The findings open up an interesting avenue for future research for designing architectures that can make better use of the aural modality.

V. CONCLUSION

We propose ViNet, a novel spatio-temporal visual-only architecture that efficiently addresses the problem of saliency detection in videos. We also explored incorporating audio for the task with AViNet by the addition of an auditory module to ViNet. We explore two different fusion techniques for combining audio-visual cues. We perform a comprehensive analysis of both models on 10 different datasets (3 visual and 7 audio-visual). Our models brings significantly gains over the state-of-the-art models. We find that audio does not seem to be playing a major role in audio-visual saliency prediction, even in models that explicitly incorporate audio. Our findings clearly illustrate the need for further explorations in this direction, leading to better models as well as curating datasets which can better utilize the auditory modality.

REFERENCES

[1] A. Tsiami, P. Koutras, and P. Maragos, "Stavis: Spatio-temporal audiovisual saliency network," in *CVPR*, 2020.

[2] N. J. Butko, L. Zhang, G. W. Cottrell, and J. R. Movellan, "Visual saliency model for robot cameras," in *ICRA*, 2008.

[3] J. F. Ferreira and J. Dias, "Attentional mechanisms for socially interactive robots–a survey," *IEEE Transactions on Autonomous Mental Development*, vol. 6, no. 2, pp. 110–125, 2014.

[4] H. Hadizadeh and I. V. Bajić, "Saliency-aware video compression," *IEEE Transactions on Image Processing*, vol. 23, no. 1, pp. 19–33, 2013.

[5] T. V. Nguyen, M. Xu, G. Gao, M. Kankanhalli, Q. Tian, and S. Yan, "Static saliency vs. dynamic saliency: a comparative study," in *ACM Multimedia*, 2013.

[6] K. B. Moorthy, M. Kumar, R. Subramanian, and V. Gandhi, "Gazed–gaze-guided cinematic editing of wide-angle monocular video recordings," in *CHI*, 2020.

[7] W. Wang, J. Shen, F. Guo, M.-M. Cheng, and A. Borji, "Revisiting video saliency: A large-scale benchmark and a new model," in *CVPR*, 2018.

[8] R. Droste, J. Jiao, and J. A. Noble, "Unified image and video saliency modeling," *arXiv preprint arXiv:2003.05477*, 2020.

[9] 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 *AAAI*, 2020.

[10] 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.

[11] K. Min and J. J. Corso, "Tased-net: Temporally-aggregating spatial encoder-decoder network for video saliency detection," in *ICCV*, 2019.

[12] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015.

[13] E. Van der Burg, C. N. Olivers, A. W. Bronkhorst, and J. Theeuwes, "Audiovisual events capture attention: Evidence from temporal order judgments," *Journal of vision*, vol. 8, no. 5, pp. 2–2, 2008.

[14] A. Coutrot, N. Guyader, G. Ionescu, and A. Caplier, "Influence of soundtrack on eye movements during video exploration," *Journal of Eye Movement Research*, vol. 5, no. 4, p. 2, 2012.

[15] 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.

[16] Y. Aytar, C. Vondrick, and A. Torralba, "Soundnet: Learning sound representations from unlabeled video," in *Advances in neural information processing systems*, 2016, pp. 892–900.

[17] M. Marszalek, I. Laptev, and C. Schmid, "Actions in context," in *CVPR*. IEEE, 2009.

[18] J. Chen, H. Song, K. Zhang, B. Liu, and Q. Liu, "Video saliency prediction using enhanced spatiotemporal alignment network," *arXiv preprint arXiv:2001.00292*, 2020.

[19] L. Jiang, M. Xu, and Z. Wang, "Predicting video saliency with object-to-motion cnn and two-layer convolutional lstm," *arXiv preprint arXiv:1709.06316*, 2017.

[20] M. Jiang, S. Huang, J. Duan, and Q. Zhao, "Salicon: Saliency in context," in *CVPR*, 2015.

[21] N. Reddy, S. Jain, P. Yarlagadda, and V. Gandhi, "Tidying deep saliency prediction architectures," *IROS*, 2020.

[22] G. Bellitto, F. P. Salanitri, S. Palazzo, F. Rundo, D. Giordano, and C. Spampinato, "Video saliency detection with domain adaption using hierarchical gradient reversal layers," *arXiv preprint arXiv:2010.01220*, 2020.

[23] W. Wang and J. Shen, "Deep visual attention prediction," *IEEE Transactions on Image Processing*, vol. 27, no. 5, pp. 2368–2378, 2017.

[24] L. Bazzani, H. Larochelle, and L. Torresani, "Recurrent mixture density network for spatiotemporal visual attention," *arXiv preprint arXiv:1603.08199*, 2016.

[25] C. Bak, A. Kocak, E. Erdem, and A. Erdem, "Spatio-temporal saliency networks for dynamic saliency prediction," *IEEE Transactions on Multimedia*, vol. 20, no. 7, pp. 1688–1698, 2017.

[26] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 5588, pp. 746–748, 1976.

[27] 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*, vol. 34, no. 5, p. 1053, 2008.

[28] A. Vatakis and C. Spence, "Crossmodal binding: Evaluating the "unity assumption" using audiovisual speech stimuli," *Perception & psychophysics*, vol. 69, no. 5, pp. 744–756, 2007.

[29] A. Coutrot and N. Guyader, "How saliency, faces, and sound influence gaze in dynamic social scenes," *Journal of vision*, vol. 14, no. 8, pp. 5–5, 2014.

[30] ——, "Multimodal saliency models for videos," in *From Human Attention to Computational Attention*. Springer, 2016, pp. 291–304.

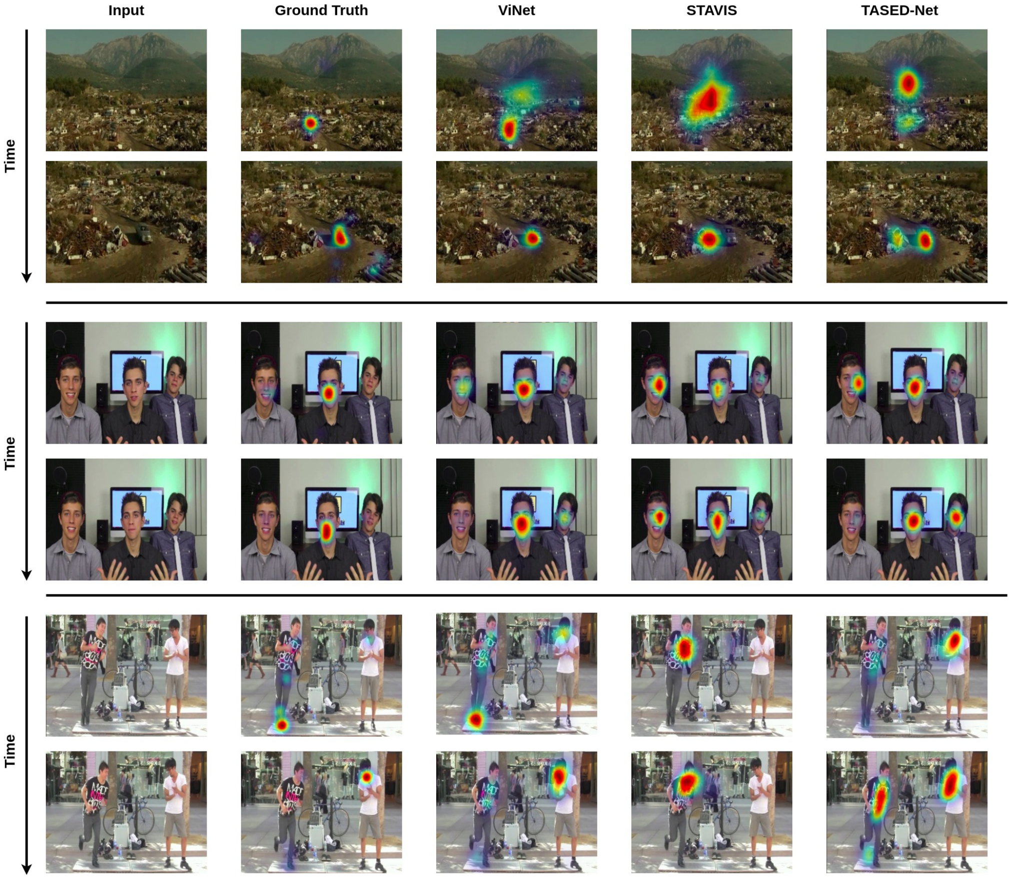

Fig. 4. Sample frames from the *Coutrot1* and *AVAD* datasets with the corresponding ground truth, ViNet, and previous state-of-the-art STAViS and TASED-Net visual saliency maps for comparisons. ViNet is able to capture the salient region in all of these 3 examples efficiently, whereas STAViS and TASED-Net are not able to capture the salient regions accurately.

[31] J. Ruesch, M. Lopes, A. Bernardino, J. Hornstein, J. Santos-Victor, and R. Pfeifer, "Multimodal saliency-based bottom-up attention a framework for the humanoid robot icub," in *ICRA*, 2008.

[32] B. Schauerte, B. Kühn, K. Kroschel, and R. Stiefelhagen, "Multimodal saliency-based attention for object-based scene analysis," in *IROS*, 2011.

[33] 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*, vol. 24, no. 11, pp. 1992–2003, 2014.

[34] X. Min, G. Zhai, K. Gu, and X. Yang, "Fixation prediction through multimodal analysis," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 13, no. 1, pp. 1–23, 2016.

[35] 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*, vol. 29, pp. 3805–3819, 2020.

[36] P. Koutras and P. Maragos, "Susinet: See, understand and summarize it," in *CVPR Workshops*, 2019.

[37] R. Arandjelovic and A. Zisserman, "Objects that sound," in *ECCV*, 2018.

[38] ——, "Look, listen and learn," in *ICCV*, 2017.

[39] A. Senocak, T.-H. Oh, J. Kim, M.-H. Yang, and I. So Kweon, "Learning to localize sound source in visual scenes," in *CVPR*, 2018.

[40] A. Owens and A. A. Efros, "Audio-visual scene analysis with self-supervised multisensory features," in *ECCV*, 2018.

[41] Y. Tian, J. Shi, B. Li, Z. Duan, and C. Xu, "Audio-visual event localization in unconstrained videos," in *ECCV*, 2018.

[42] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy, "Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification," in *ECCV*, 2018.

[43] W. Wang, J. Shen, J. Xie, M.-M. Cheng, H. Ling, and A. Borji, "Revisiting video saliency prediction in the deep learning era," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 1, pp. 220–237, 2019.

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

[45] M. D. Rodriguez, J. Ahmed, and M. Shah, "Action mach a spatio-temporal maximum average correlation height filter for action recognition," in *CVPR*, 2008.

[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*, vol. 3, no. 1, pp. 5–24, 2011.

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

[48] M. Gygli, H. Grabner, H. Riemenschneider, and L. Van Gool, "Creating summaries from user videos," in *ECCV*, 2014.