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

Samyak Jain\textsuperscript{1}, Pradeep Yarlagadda\textsuperscript{1}, Shreyank Jyoti\textsuperscript{1}, Shyamgopal Karthik\textsuperscript{1}, Ramanathan Subramanian\textsuperscript{2}, Vineet Gandhi\textsuperscript{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 model 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 \cite{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 \url{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 \cite{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 \cite{3,4,5,6}, robotic camera control \cite{2}, motion tracking \cite{7}, stream compression \cite{8,9}, video captioning \cite{10}, automated cinematic editing \cite{11}, etc.

Video SP models primarily employ visual information to predict gaze. Larger datasets like DHF1K \cite{12} discard audio during ground truth collection, and ask users to look at silent videos. End-to-end deep saliency models are then trained

\textsuperscript{1} CVIT, KCIS, International Institute for Information Technology, Hyderabad samyak.j@research.iiit.ac.in
\textsuperscript{2}University of Canberra
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 [18]. Coutrot et al. [19] 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 [19]. Consequently, recent efforts have explored multi-modal (audio-visual) video SP [20], [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 [21]. We explore two fusion strategies, similar to [1], [20] 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 [20], [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 [22] and DHF1K [12] 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. [12] combine frame-level image features using a ConvLSTM. SalEMA [15] 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 [14] adds further complexity to basic ConvLSTM architecture through a shuffled multi-level attention module and a frame correlation module. STRA-Net [23] learns an alignment module, and then aligned frames are sent into a Bi-ConvLSTM. [24] propose a novel construction of LSTM (2C-LSTM) with two sub-networks to focus on objectness and motion, respectively. UNISAL [13] 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 [13], [12] borrow and extend ideas (hierarchical features, transfer learning, multi-branch architectures, etc.) from the models trained for static image saliency prediction [25], [26].

3D convolutional architectures have also been explored for the task. These methods typically rely on action detection networks as their backbone. TASED-Net [16] uses SSD as an encoder to extract spatial features while jointly aggregating all the temporal information in order to produce a
Fig. 3. AViNet Architecture overview. Removing the audio branch, the resulting architecture is ViNet.

The architecture uses the S3D network [47] 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. $X_1, X_2$ 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 \(x_{\text{clip}} \in R^{3 \times T_0 \times H_0 \times W_0}\), where \(T_0 = 32\). It generates a lower-resolution activation map \(X_4 \in R^{C \times T \times H \times W}\), where \(C = 1024\) and \(T, H, W = \frac{T_0}{8} \times \frac{H_0}{32} \times \frac{W_0}{32}\).

For audio representation, we employ SoundNet [21], 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, \(y_{\text{audio}} \in R^{1 \times T' \times 1}\) 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], [20], we explore two types of fusion techniques. First is a simple concatenation of encoded audio and video features which was used in [20]. We repeat the audio features to match the dimensions of visual features and combine them across the channel dimension. Then we apply \(1 \times 1\) 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 \(x_1 \in R^{1024 \times z_0}\). Similarly, the audio features are collapsed as a vector \(x_2 \in R^{1024 \times y_0}\). The bilinear fusion is defined as

\[
y = x_1^T A x_2 + b
\]

where \(A \in R^{z_0 \times z \times y_0}\) and \(b \in R^{z \times 1}\) 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 \(F_1, ..., F_{t+1}\) 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 [50]. We carry out the tests and comparisons on these three datasets.

DHF1K [12] 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 [22] 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 [50] 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 [51] 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 [34], [35] 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 [39] contains 45 short clips of 5-10 sec duration with several audio-visual scenes, e.g., dancing, guitar playing, birds singing, etc.

ETMD dataset [52] contains 12 videos from six different Hollywood movies.

SumMe dataset [53] 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.
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. FPS values are taken from the leaderboard and are computed on different hardware, hence, a direct comparison cannot be made. ViNet is evaluated on a GTX 1080 Ti GPU and runs at a speed adequate for most real-time applications.

| Dataset | FPS | DHF1K  |好莱坞2| UCF-Sports |
|---------|-----|--------|--------|------------|
|         | CC  | sAUC   | AUC    | NSS        | SIM        |
|         | 0.613 | 0.708 | 0.919 | 3.18 | 0.487 |
|         | 0.5178 | 0.5212 | 0.5212 | 0.5212 | 0.5212 |

**A. Experimental Setup**
For training ViNet, clips with $T$ consecutive frames were randomly selected from the dataset. Each frame is resized to $224 \times 384$ and trained with a batch size of 8. The optimizer used is Adam, and the learning rate is set to be $1 \times 10^{-4}$. The network is initialized 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.

**B. Evaluation Metrics**
We evaluate our method on five standard evaluations metrics [54]. The CC metric computes the Pearson’s Correlation between the ground truth and the predicted maps. The Similarity metric (SIM) computes the histogram intersection. The Area Under the ROC Curve (AUC) treats saliency map as a binary classifier of fixations at various threshold values, and an ROC curve is swept out by measuring the true and false positive rates under each binary classifier. sAUC is a variation of AUC where negatives are sampled from fixation locations from other images. NSS is computed as the average normalized saliency at fixated locations.

**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:

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

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

**D. Comparison with state-of-the-art**

- **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). We

---

1 The challenge website can be found here: https://mmcheng.net/videosal/
TABLE IV
COMPARISON RESULTS ON THE DIEM, Coutrot1, Coutrot2, AVAD, ETMD AND SumMe TEST SETS.

| Model Name       | CC sAUC | AUC   | NSS   | SIM   | CC sAUC | AUC   | NSS   | SIM   | CC sAUC | AUC   | NSS   | SIM   |
|------------------|---------|-------|-------|-------|---------|-------|-------|-------|---------|-------|-------|-------|
| ACLNet [48]      | 0.527   | 0.622 | 0.869 | 2.02  | 0.427   | 0.542 | 0.85  | 1.92  | 0.361   | 0.448 | 0.594 | 0.926 | 3.16   |
| TASED-Net [16]   | 0.557   | 0.657 | 0.818 | 2.16  | 0.461   | 0.479 | 0.867 | 2.18  | 0.388   | 0.437 | 0.611 | 0.921 | 3.17   |
| STAVIS [1]       | 0.579   | 0.674 | 0.883 | 2.26  | 0.482   | 0.472 | 0.868 | 2.11  | 0.393   | 0.734 | 0.71  | 0.958 | 5.28   |
| ViNet (NF)       | 0.371   | 0.695 | 0.886 | 2.28  | 0.468   | 0.390 | 0.873 | 2.46  | 0.406   | 0.645 | 0.72  | 0.949 | 5.11   |
| ViNet            | 0.626   | 0.723 | 0.898 | 2.47  | 0.483   | 0.551 | 0.886 | 2.68  | 0.423   | 0.724 | 0.739 | 0.95  | 5.61   |
| AVINet (B)       | 0.632   | 0.719 | 0.899 | 2.53  | 0.498   | 0.56  | 0.889 | 2.73  | 0.425   | 0.754 | 0.742 | 0.951 | 5.95   |
| AVINet (C)       | 0.631   | 0.720 | 0.897 | 2.50  | 0.497   | 0.556 | 0.887 | 2.68  | 0.426   | 0.753 | 0.743 | 0.951 | 5.81   |

show a qualitative 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. We also present results on ViNet(NF) baseline model, which is a trained of DHFIK 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.

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 DHFIK 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], [20], 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 [20]. 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 V ViNet and AVINet give fairly similar performance across all three
TABLE VI
CC AND SIM METRICS 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.

| Method      | CC     | SIM    |
|-------------|--------|--------|
| AVINet(B)   | 0.9977 | 0.9979 |
| AVINet(C)   | 0.9978 | 0.9990 |
| STAVIS      | 0.9980 | 0.9981 |

categories, giving solid gains over other methods.

G. 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). 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 observations contrast with cognitive studies, which suggest clear differences in human gaze patterns when the videos are watched with or without audio [19]. 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] G. Schillaci, S. Bodiroza, and V. V. Hafner, “Evaluating the effect of saliency detection and attention manipulation in human-robot interaction,” International Journal of Social Robotics, vol. 5, no. 1, pp. 139–152, 2013.
[5] Z. Chang, J. Matias Di Martino, Q. Qiu, S. Espinosa, and G. Sapiro, “Salgaze: Personalizing gaze estimation using visual saliency,” in ICCV Workshops, 2019.
[6] V. Mavani, S. Raman, and K. P. Myiyapram, “Facial expression recognition using visual saliency and deep learning,” in ICCV workshops, 2017.
[7] G. Zhang, Z. Yuan, N. Zheng, X. Sheng, and T. Liu, “Visual saliency based object tracking,” in Asian conference on computer vision. Springer, 2009, pp. 193–203.
[8] H. Hadizadeh and I. V. Bajic, “Saliency-aware video compression,” IEEE Transactions on Image Processing, vol. 23, no. 1, pp. 19–33, 2014.
[9] R. Gupta, M. T. Khanna, and S. Chaudhury, “Visual saliency guided video compression algorithm,” Signal Processing: Image Communication, vol. 28, no. 9, pp. 1006–1022, 2013.
[10] 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.
[11] K. B. Moorthy, M. Kumar, R. Subramanian, and V. Gandhi, “Gazed–gaze-guided cinematic editing of wide-angle monocular video recordings,” in CHI, 2020.
[12] 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.
[13] R. Droste, J. Jiao, and J. A. Noble, “Unified image and video saliency modeling,” arXiv preprint arXiv:2003.05477, 2020.
[14] 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.
[15] 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.
[16] K. Min and J. J. Corso, “Tased-net: Temporally-aggregating spatial encoder-decoder network for video saliency detection,” in ICCV, 2019.
[17] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in MICCAI, 2015.
[18] 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.
[19] 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.
[20] R. H. Tavakoli, A. Borji, E. Rahtu, and J. Kannala, “Dave: A deep audio-visual embedding for dynamic saliency prediction,” arXiv preprint arXiv:1905.10693, 2019.
[21] Y. Aytar, C. Vondrick, and A. Torralba, “Soundnet: Learning sound representations from unlabeled video,” in NeurIPS, 2016.
[22] M. Marszalek, I. Laptev, and C. Schmid, “Actions in context,” in CVPR, 2009.
[23] 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.
[24] 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.
[25] M. Jiang, S. Huang, J. Duan, and Q. Zhao, “Salicon: Saliency in context,” in CVPR, 2015.
[26] N. Reddy, S. Jain, P. Yarlagadda, and V. Gandhi, “Tidying deep saliency prediction architectures,” IROS, 2020.
[27] G. Bellitto, F. P. Salanitri, S. Palazzo, F. Rundo, D. Giordano, and C. Spampinato, “Video saliency detection with domain adaptation using hierarchical gradient reversal layers,” arXiv preprint arXiv:2010.01220, 2020.
Fig. 4. Sample frames from the Coutrot1 and AVAD datasets with the corresponding ground truth, ViNet, and previous state-of-the-art STA ViS and TASED-Net visual saliency maps for comparisons. ViNet is able to capture the salient region in all of these 3 examples efficiently, whereas STA ViS and TASED-Net are not able to capture the salient regions accurately. Complete video for all figures is available in the supplementary material.

[29] L. Bazzani, H. Larochelle, and L. Torresani, “Recurrent mixture density network for spatiotemporal visual attention,” arXiv preprint arXiv:1603.08199, 2016.

[30] 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.

[31] H. McGurk and J. MacDonald, “Hearing lips and seeing voices,” Nature, vol. 264, no. 5588, pp. 746–748, 1976.

[32] 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.

[33] 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.

[34] 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.

[35] ——, “Multimodal saliency models for videos,” in From Human Attention to Computational Attention. Springer, 2016, pp. 291–304.

[36] 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.

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

[38] 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.

[39] 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.

[40] 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.

[41] P. Koutras and P. Maragos, “Susinet: See, understand and summarize it,” in CVPR Workshops, 2019.

[42] R. Arandjelovic and A. Zisserman, “Objects that sound,” in ECCV, 2018.

[43] ——, “Look, listen and learn,” in ICCV, 2017.

[44] 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.

[45] A. Owens and A. A. Efros, “Audio-visual scene analysis with self-supervised multisensory features,” in ECCV, 2018.

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

[47] 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.

[48] 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.

[49] 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.

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

[51] 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.

[52] 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.

[53] M. Gygli, H. Grabner, H. Riemenschneider, and L. Van Gool, “Creating summaries from user videos,” in ECCV, 2014.

[54] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, “What do different evaluation metrics tell us about saliency models?” IEEE transactions on pattern analysis and machine intelligence, vol. 41, no. 3, pp. 740–757, 2018.