ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-6874-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICASSP49660.2025.10888852

Minimalistic Video Saliency Prediction via Efficient Decoder & Spatio Temporal Action Cues

Rohit Girmaji, Siddharth Jain, Bhav Beri, Sarthak Bansal, Vineet Gandhi
CVIT, IIIT Hyderabad, India
{rohit.girmaji, siddharth.jain, bhav.beri}@research.iiit.ac.in, sarthak.bansal@students.iiit.ac.in, vgandhi@iiit.ac.in

***Abstract*—This paper introduces ViNet-S, a 36MB model based on the ViNet architecture with a U-Net design, featuring a lightweight decoder that significantly reduces model size and parameters without compromising performance. Additionally, ViNet-A (148MB) incorporates spatio-temporal action localization (STAL) features, differing from traditional video saliency models that use action classification backbones. Our studies show that an ensemble of ViNet-S and ViNet-A, by averaging predicted saliency maps, achieves state-of-the-art performance on three visual-only and six audio-visual saliency datasets, outperforming transformer-based models in both parameter efficiency and real-time performance, with ViNet-S reaching over 1000fps.**

***Index Terms*—Video Saliency Prediction, Efficient Deep Learning, Spatio Temporal Action Cues**

I. Introduction

Human visual attention (HVA) enables selective focus on relevant stimuli, a capability that computational saliency prediction (SP) aims to replicate in dynamic scenes. The formal approach to addressing this task involves initially recording human gaze using an eye-tracking hardware device and subsequently employing this data as the reference point for training predictive models. SP models have made substantial progress over the years and have shown considerable benefits across a wide range of applications such as intelligent robotic behaviour [1], automated cinematic editing [2], human-computer interaction [3]–[6], and autonomous driving [7].

In the deep learning era, early SP methods used two-stream approaches [8], [9] or recurrent networks [10], [11], which struggled with long-range dependencies and spatial-temporal cues. 3D convolution-based model [12], [13] architectures then followed, which typically utilize action classification backbones like S3D [14] pre-trained on the Kinetics dataset [15]. ViNet [12], a fully convolutional encoder-decoder, uses hierarchical features with UNet-like [16] skip connections. STSANet [17] employs spatio-temporal self-attention but is too large for practical use. Recent approaches like TMFI-Net [18] and THTD-Net [19] use Video Swin Transformer for saliency prediction, focusing on long-range temporal dependencies.

Prior works have also explored combining audio and visual modalities for saliency prediction. STAViS [20] combines spatio-temporal visual and auditory features with linear weighting. TSFP-Net [21] builds a temporal-spatial feature pyramid, fusing audio and visual features with attention mechanisms. VAM-Net [22], VASM [23] employs multi-stream and multi-modal networks to predict saliency maps. CASP-Net [24] associates video frames with sound sources using a two-stream encoder. Recently, DiffSal [25] introduced a diffusion-based approach for audio-visual saliency modelling; however, it suffers from heightened computational complexity and substantially slower inference speeds. In contrast, Our work focuses solely on optimizing the visual modality.

We revisit 3D convolutions with the ViNet architecture [12], proposing ViNet-S, a computationally efficient model with a lightweight decoder using filter groups [26] and channel shuffle layers [27], achieving a threefold reduction in size and parameters while improving SP performance. We also identify limitations in using action classification backbones like S3D [14], which may miss background actions due to a focus on primary motion. Instead, we propose ViNet-A, leveraging Spatio-Temporal Action Localization (STAL) [28], [29] with our lightweight decoder, which localizes and classifies actions within the scene, better capturing scene essence. ViNet-A excels, particularly in human-centric datasets like MVVA [23], by focusing on the most relevant features, such as the salient face in group settings.

We further introduce ViNet-E, an ensemble of ViNet-S and ViNet-A, combining their strengths by averaging their predicted saliency maps. Despite its compact design, ViNet-E outperforms transformer-based approaches on various datasets without using audio cues. Our contributions include: 1) ViNet-S: A lightweight model with 9 million parameters, surpassing the original ViNet [12] in performance. 2) ViNet-A: Utilizing a STAL backbone for enhanced performance in videos with multiple subjects. 3) ViNet-E: An ensemble of ViNet-S and ViNet-A, achieving SOTA results across multiple datasets. 4) Extensive experiments on nine datasets, providing qualitative and quantitative insights.

II. Proposed Model Architecture

We propose an end-to-end trainable visual-only model called ViNet-A (Figure 1). It is a fully 3D-convolutional encoder-decoder architecture consisting of a SlowFast network [29] as the video encoder, a convolutional neck, and an efficient, lightweight decoder to reduce computational costs for predicting the saliency map. We also propose a variation of the ViNet architecture [12], ViNet-S, which utilizes our efficient decoder, resulting in a small model while surpassing the original ViNet's performance. Lastly, we propose an ensemble

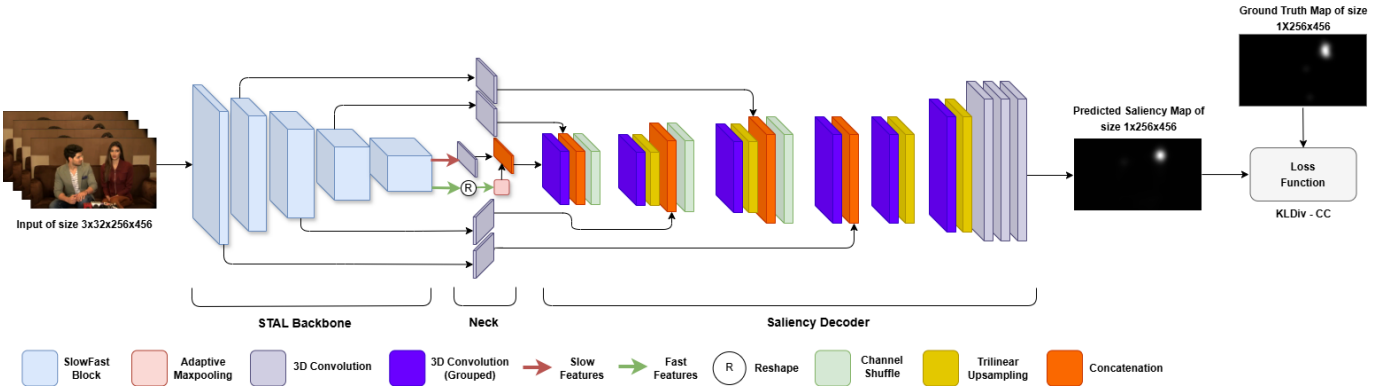

Fig. 1: Our Model (ViNet-A) Architecture for SP (Best viewed in colour)

of the two proposed models, ViNet-E. We elaborate on the proposed models in the following sections.

A. ViNet-A

1) Backbone: Our model utilizes the SlowFast network [29], pre-trained on the AVA actions dataset [30] as its video encoder. This backbone effectively captures localized actions across spatial and temporal dimensions. The SlowFast network comprises of two parallel pathways: the Slow pathway, which captures spatial semantics at a low frame rate, and the Fast pathway, which focuses on fine-grained temporal motion at a high frame rate. Both pathways are 3D convolutional networks that combine information through lateral connections, which are subsequently used as skip connections in the saliency decoder.

2) Neck: The neck uses 1×1 convolutional blocks to reduce the number of channels, lowering computational overhead. We reduce the number of channels in X_{slow} by half. X_{fast} is reshaped to double its channels while halving its temporal dimension and then passed through an adaptive max pool to align its temporal dimension with X_{slow}. The two are concatenated channel-wise, resulting in fused SlowFast features, $X_{slowfast} \in \mathbb{R}^{1536\times8\times16\times29}$. Similarly, hierarchical features X_1, X_2, X_3, and X_4 are processed through 1×1 convolutional blocks to halve their channels, improving computational efficiency.

3) Saliency Decoder: The Saliency Decoder consists of six decoding blocks with 3D convolutions using filter groups [26], trilinear upsampling and channel shuffle [27] layers to reduce computational costs while preserving accuracy. SlowFast features, $X_{slowfast}$, are fed into the decoder, with hierarchical features X_i passed as skip connections. All 3D convolutions, except the last block, utilize filter groups with 32, 16, 8, 8, 4 and 2 groups, respectively, with channel shuffle layers applied after the first three grouped convolutions. We experimented with different filter groups and channel shuffle layer configurations and found this setup optimal. ReLU activations follow every convolutional layer, except for the last, which uses Sigmoid to predict the saliency map.

B. ViNet-S & ViNet-E

ViNet-S employs the S3D [14] backbone as its video encoder and the lightweight decoder with grouped convolutions and channel shuffle layers, similar to the ViNet-A saliency decoder described above.

ViNet-E is an ensemble of the proposed models, ViNet-S and ViNet-A, which generates a saliency map by performing a simple pixel-wise mean of the two predicted saliency maps. Since both models predict saliency maps of different sizes, the ViNet-S prediction is upsampled to match ViNet-A before averaging.

III. Experiments

a) Datasets: We conduct experiments on three visual-only saliency datasets - DHF1K [11], Hollywood-2 [31], and UCF-Sports [31] and six audio-visual saliency prediction datasets- AVAD [32], Coutrot1 [33], [34], Coutrot2 [35], DIEM [36], ETMD [37] and MVVA [23].

b) Training: Following [12], we input a clip of 32 consecutive frames to the ViNet-S model and use the ground truth saliency map of the 32$^{\text{nd}}$ frame for supervision and prediction. For the ViNet-A model, the input consists of 32 frames sampled from a window of 64 consecutive frames by selecting every alternate frame. We use the ground truth saliency map of the 33$^{\text{rd}}$ frame for supervision and prediction, akin to action label predictions in STAL models [28]. Both models are trained using the Adam optimizer with a learning rate of 10^{-4} and batch size of 8 for ViNet-S and 6 for ViNet-A.

For evaluating our model on DHF1K, we use the validation set due to unavailable annotations for the test set, as in prior efforts [38], [39]. We use the standard train and test sets provided for training on datasets Hollywood-2, UCF-Sports and DIEM. For Coutrot1, Coutrot2, AVAD, and ETMD, we perform 3-fold cross-validation and report average metrics across the splits. For MVVA, we follow [22] and perform training on a random split.

c) Evaluation Metrics: We evaluate our method on five standard evaluation metrics whose details can be found in

TABLE I: Quantitative comparison of model sizes and performance on visual-only datasets.

(a) Results on DHF1K validation set, UCF-Sports and Hollywood2 test sets. Best results highlighted in red and second best in blue.

METHOD	DHF1K (Validation Set)				UCF-Sports				Hollywood2			
	CC↑	NSS↑	AUC-J↑	SIM↑	CC↑	NSS↑	AUC-J↑	SIM↑	CC↑	NSS↑	AUC-J↑	SIM↑
ACLNet [11]	0.434	2.35	0.890	0.315	0.510	2.56	0.897	0.406	0.623	3.08	0.913	0.542
TASED-Net [13]	0.481	2.706	0.894	0.362	0.582	2.920	0.899	0.469	0.646	3.302	0.918	0.507
UNISAL [10]	0.490	2.77	0.901	0.390	0.644	3.38	0.918	0.523	0.673	3.90	0.934	0.542
ViNet [12]	0.521	2.957	0.919	0.388	0.673	3.620	0.924	0.522	0.693	3.730	0.930	0.550
TSFP-Net [21]	0.529	3.009	0.919	0.397	0.685	3.698	0.923	0.561	0.711	3.910	0.936	0.571
STSANet [17]	0.539	3.082	0.920	0.411	0.721	3.927	0.936	0.560	0.705	3.908	0.938	0.579
TMFI-Net [18]	0.554	3.201	0.924	0.428	0.707	3.863	0.936	0.565	0.739	4.095	0.940	0.607
THTD-Net [19]	0.553	3.188	0.924	0.425	0.711	3.840	0.933	0.565	0.726	3.965	0.939	0.585
DiffSal [25]	0.533	3.066	0.918	0.405	0.685	3.483	0.928	0.543	0.765	3.955	0.951	0.610
ViNet-S	0.529	3.008	0.919	0.399	0.673	3.652	0.930	0.530	0.728	3.941	0.941	0.582
ViNet-A	0.525	3.019	0.916	0.399	0.734	4.108	0.940	0.586	0.756	4.119	0.945	0.604
ViNet-E	0.549	3.134	0.922	0.409	0.744	4.156	0.941	0.587	0.766	4.168	0.947	0.609

(b) Quantitative comparison of model sizes & parameters

Model	Size (MB)	# Params (Million)
ACLNet [11]	250	65.54
TASED-Net [13]	82	21.5
STAViS [20]	79.19	20.76
UNISAL [10]	15.5	4.06
ViNet [12]	124	32.5
TSFP-Net [21]	58.4	15.3
STSA-Net [17]	643	168.56
TMFI-Net [18]	234	61.34
THTD-Net [19]	220	57.67
CASP-Net [24]	196.91	51.62
DiffSal [25]	269	70.54
ViNet-S	36.24	9.5
ViNet-A	147.6	38.69
ViNet-E	183.84	48.19

TABLE II: Quantitative comparison results on the AVAD, Coutrot1, Coutrot2 and ETMD test sets.

METHOD	Coutrot1				Coutrot2				ETMD				AVAD			
	CC↑	NSS↑	AUC-J↑	SIM↑	CC↑	NSS↑	AUC-J↑	SIM↑	CC↑	NSS↑	AUC-J↑	SIM↑	CC↑	NSS↑	AUC-J↑	SIM↑
ACLNet [11]	0.425	1.92	0.85	0.361	0.448	3.16	0.926	0.322	0.477	2.36	0.915	0.329	0.580	3.17	0.905	0.446
TASED-Net [13]	0.479	2.18	0.867	0.388	0.437	3.17	0.921	0.314	0.509	2.63	0.916	0.366	0.601	3.16	0.914	0.439
STAViS [20]	0.458	1.99	0.861	0.384	0.652	4.19	0.940	0.447	0.560	2.84	0.929	0.412	0.604	3.07	0.915	0.443
ViNet [12]	0.551	2.68	0.886	0.423	0.724	5.61	0.95	0.466	0.569	3.06	0.928	0.409	0.694	3.82	0.928	0.504
TSFP-Net [21]	0.57	2.75	0.894	0.451	0.718	5.30	0.957	0.516	0.576	3.09	0.932	0.433	0.688	3.79	0.932	0.530
CASP-Net [24]	0.561	2.65	0.889	0.456	0.788	6.34	0.963	0.585	0.620	3.34	0.940	0.478	0.691	3.81	0.933	0.528
ViNet-S	0.574	2.876	0.898	0.449	0.754	6.103	0.958	0.547	0.599	3.268	0.941	0.458	0.712	4.090	0.935	0.540
ViNet-A	0.600	3.033	0.900	0.459	0.862	6.8	0.961	0.638	0.623	3.379	0.941	0.458	0.709	4.094	0.933	0.534
ViNet-E	0.614	3.085	0.905	0.465	0.854	6.762	0.962	0.628	0.632	3.437	0.943	0.468	0.729	4.167	0.938	0.547

TABLE III: Quantitative comparison results on the DIEM and MVVA test sets.

METHOD	DIEM			
	CC↑	NSS↑	AUC-J↑	SIM↑
ACLNet [11]	0.522	2.02	0.869	0.427
TASED-Net [13]	0.557	2.16	0.881	0.461
STAViS [20]	0.579	2.26	0.883	0.482
ViNet [12]	0.626	2.47	0.898	0.483
TSFP-Net [21]	0.651	2.62	0.906	0.527
CASP-Net [24]	0.655	2.61	0.906	0.543
ViNet-S	0.673	2.732	0.908	0.533
ViNet-A	0.675	2.742	0.908	0.547
ViNet-E	0.701	2.840	0.913	0.566

METHOD	MVVA			
	CC↑	NSS↑	AUC-J↑	KLDIV↓
VASM [23]	0.722	3.976	0.905	0.823
VAM-Net [22]	0.741	4.002	0.912	0.783
TASED-Net [13]	0.653	3.319	0.905	0.970
STAViS [20]	0.77	3.060	0.91	0.80
ViNet [12]	0.81	4.470	0.93	0.75
ViNet-S	0.802	4.617	0.933	0.715
ViNet-A	0.825	4.823	0.934	0.678
ViNet-E	0.828	4.816	0.936	0.663

[40]: AUC-Judd (AUC-J), Similarity Metric (SIM), Correlation Coefficient(CC), Normalized Scanpath Saliency(NSS) and Kullback-Leibler Divergence(KLDiv). Except for KLDiv, higher metric values indicate better model performance.

d) Loss Function: We utilize a combination of the above evaluation metrics, a standard technique in saliency tasks [40]. Through experimentation with different combinations, we found that the optimal results for most datasets were achieved with the loss function: $Loss = KLDiv(P, Q) - CC(P, Q)$, where P & Q represent the predicted saliency map and ground truth, respectively.

IV. Results and Discussions

We evaluate the proposed models by comparing them against thirteen different methods from previous research. These include four 3D convolution-based approaches: ViNet [12], TASED-Net [13], STAVIS [20], and TSFP-Net [21]; two methods utilizing recurrent networks: ACLNet [11] and UNISAL [10]; four models employing transformers: STSANet [17], THTD-Net [19], CASP-Net [24], TMFI-Net [18]; one diffusion-based model: DiffSal [25] and a couple of multi-branch network methods: VAM-Net [22] and VASM [23]. Six of these models (STAVIS, CASP-Net, TSFP-Net, VAM-Net, DiffSal and VASM) additionally employ audio information in their approach. We report results directly from the corresponding papers when available. If the code is publicly available and executable, we compute their results on other datasets.

a) Visual Only Datasets: Table Ia presents results on the visual-only datasets. The model sizes and the number of parameters of the studied models are presented in Table Ib. We observe that ViNet-E achieves the best performance on UCF-Sports and Hollywood2 datasets [31], while achieving competent results on the DHF1K dataset [11]. Interestingly, ViNet-A also outperforms the previous methods on the UCF-Sports and Hollywood2 datasets. Its strong performance on these two human-centric datasets clearly demonstrates the advantages of using an STAL backbone over an action classification backbone. Notably, all three proposed models, including ViNet-S, consistently surpass the base ViNet model.

The ViNet-S model recovers most of the underlying performance in all the cases while using only a tiny fraction of

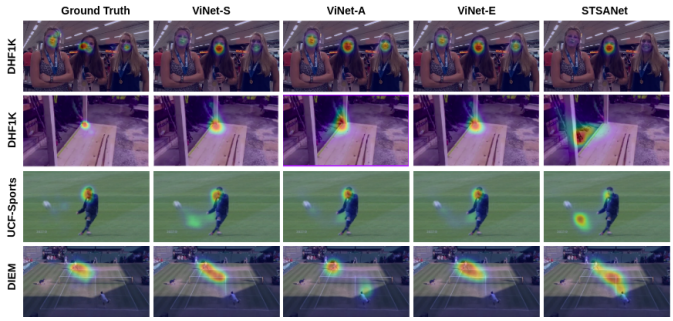

Fig. 2: Qualitative results: Comparing Ground Truth with the predicted saliency maps of our models and STSANet on three different datasets - DHF1K, UCF-Sports and DIEM.

the parameters. For instance, on the Hollywood2 dataset, the largest SP dataset with 884 videos in the test set, ViNet-S recovers over 98.5% performance on the CC metric compared to the transformer-based SOTA TMFI-Net, while bringing over six-fold reduction in terms of number of parameters (Table I). Interestingly, ViNet-S outperforms TMFI-Net on the AUC-J metric. UNISAL is the only model lighter than the ViNet-S model. However, it consistently underperforms in comparison, possibly due to its recurrent architecture.

b) Audio Visual Datasets: Table II and Table III present results on the audio-visual datasets. The proposed ViNet-S, ViNet-A, and ViNet-E consistently outperform prior models across all six datasets, consistently ranking among the top two models. The videos in the Coutrot2 and MVVA datasets emphasize multi-person interactions. Notably, ViNet-A achieves significant improvements on both datasets, maintaining a consistent performance trend with the other human-centric datasets. On MVVA (the largest audio-visual saliency dataset), while only using the visual modality ViNet-A brings over 20% gains on NSS metric over the complex multi-branch VAM-Net, which uses an explicit combination of motion, texture, face and audio features.

On four out of the six audio-visual datasets, i.e. DIEM, AVAD, Coutrot1, and MVVA, the smaller ViNet-S surpasses all the previous methods. The consistent performance improvements of the ViNet-E model validate the effectiveness of the proposed ensemble strategy, establishing a new SOTA in most datasets. Another notable observation is that incorporating audio information does not appear to provide a significant advantage for the task of SP. Consistent with prior studies [12], [21], [41], we found that several audio-visual models [20], [25], in reality, are not exploiting the audio information. At inference, the models appear agnostic to the audio information, i.e., the results remain the same irrespective of sending the random audio or zero audio. This represents a significant scientific flaw that requires further investigation in future research, and comparisons with their results should be approached with caution. Although ViNet-E outperforms their audio-visual version on several datasets, we limit our comparisons only to their visual only model.

c) Qualitative comparisons: Figure 2 shows the qualitative performance of our ViNet-S, ViNet-A and ViNet-E models on video sequences from three different datasets: DHF1K, UCF-Sports and DIEM. We observe that STAL features efficiently capture the interaction between an actor/object with the context (surrounding) as evident in the strong performance of our model. ViNet-E is consistently closer to the ground truth in different settings than all other models, including STSANet.

d) Computational load: Table Ib compares the different models in terms of models size and number of parameters. The proposed decoder significantly reduces the number of parameters in ViNet-S compared to the original ViNet model. Aside from UNISAL, ViNet-S is the most efficient in terms of model size and parameters among the compared models. We observe that switching from the S3D backbone in ViNet-S to the SlowFast backbone in ViNet-A leads to significant parameter gains. Notably, ViNet-A's decoder contains only 1.6 million parameters, while the SlowFast backbone accounts for the remaining 37 million. Lastly, the ViNet-E model remains smaller than state-of-the-art transformer-based models (e.g., TMFI-Net and THTD-Net) in both model size and parameter count.

The non-autoregressive design of the proposed ViNet models enables parallel processing, providing a significant advantage over autoregressive models such as UNISAL, which rely on frame-level recurrence. On an Nvidia RTX 4090 GPU, ViNet-S, ViNet-A, and ViNet-E models achieve runtimes of approximately 200fps, 120fps, and 90fps, respectively, in a real-time processing setup (with a batch size of one). With a batch size of eight, ViNet-S reaches an impressive 1070fps.

V. Conclusion

This work introduces two efficient models, ViNet-S and ViNet-A, characterized by their simple architectural design choices. ViNet-S is lightweight yet matches or surpasses most convolutional methods, while ViNet-A, which utilizes localized action features, consistently performs well on human-centric datasets with multiple subjects. ViNet-E, the ensemble model, leverages the complementary nature of action classification and detection to achieve state-of-the-art results on both visual and audio-visual datasets, even without audio cues. Using pixel-wise averaging enhances performance, suggesting new avenues for integrating global and localized action features. While this study emphasizes model optimization primarily through architectural refinements, future work would aim to investigate and integrate ideas from model compression and knowledge distillation methodologies.

References

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

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

[3] Z. Chang, J. Matias Di Martino, Q. Qiu, S. Espinosa, and G. Sapiro, "Salgaze: Personalizing gaze estimation using visual saliency," in *International Conference on Computer Vision Workshops (ICCVW)*, 2019.

[4] 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.

[5] V. Mavani, S. Raman, and K. P. Miyapuram, "Facial expression recognition using visual saliency and deep learning," in *International Conference on Computer Vision Workshops (ICCVW)*, 2017.

[6] G. Schillaci, S. Bodiroža, and V. V. Hafner, "Evaluating the effect of saliency detection and attention manipulation in human-robot interaction," *International Journal of Social Robotics*, vol. 5, pp. 139–152, 2013.

[7] F. Lateef, M. Kas, and Y. Ruichek, "Saliency heat-map as visual attention for autonomous driving using generative adversarial network (gan)," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 6, pp. 5360–5373, 2021.

[8] A. Kocak, E. Erdem, and A. Erdem, "A gated fusion network for dynamic saliency prediction," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 14, no. 3, pp. 995–1008, 2021.

[9] K. Zhang and Z. Chen, "Video saliency prediction based on spatial-temporal two-stream network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 12, pp. 3544–3557, 2019.

[10] R. Droste, J. Jiao, and J. A. Noble, "Unified image and video saliency modeling," in *European Conference on Computer Vision (ECCV)*. Springer, 2020, pp. 419–435.

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

[12] S. Jain, P. Yarlagadda, S. Jyoti, S. Karthik, R. Subramanian, and V. Gandhi, "Vinet: Pushing the limits of visual modality for audio-visual saliency prediction," in *IROS*. IEEE, 2021, pp. 3520–3527.

[13] K. Min and J. J. Corso, "Tased-net: Temporally-aggregating spatial encoder-decoder network for video saliency detection," in *International Conference on Computer Vision (ICCV)*, 2019, pp. 2394–2403.

[14] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy, "Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification," in *European Conference on Computer Vision (ECCV)*, 2018.

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

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

[17] Z. Wang, Z. Liu, G. Li, Y. Wang, T. Zhang, L. Xu, and J. Wang, "Spatio-temporal self-attention network for video saliency prediction," *IEEE Transactions on Multimedia*, vol. 25, pp. 1161–1174, 2021.

[18] X. Zhou, S. Wu, R. Shi, B. Zheng, S. Wang, H. Yin, J. Zhang, and C. Yan, "Transformer-based multi-scale feature integration network for video saliency prediction," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 12, pp. 7696–7707, 2023.

[19] M. Moradi, S. Palazzo, and C. Spampinato, "Transformer-based video saliency prediction with high temporal dimension decoding," *VISIGRAPP*, 2024.

[20] A. Tsiami, P. Koutras, and P. Maragos, "Stavis: Spatio-temporal audio-visual saliency network," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 4766–4776.

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

[22] M. Qiao, Y. Liu, M. Xu, X. Deng, B. Li, W. Hu, and A. Borji, "Joint learning of audio-visual saliency prediction and sound source localization on multi-face videos," *International Journal of Computer Vision (IJCV)*, vol. 132, pp. 2003–2025, 2023.

[23] Y. Liu, M. Qiao, M. Xu, B. Li, W. Hu, and A. Borji, "Learning to predict salient faces: A novel visual-audio saliency model," in *European Conference on Computer Vision (ECCV)*, 2020, pp. 413–429.

[24] J. Xiong, G. Wang, P. Zhang, W. Huang, Y. Zha, and G. Zhai, "Casp-net: Rethinking video saliency prediction from an audio-visual consistency perceptual perspective," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 6441–6450.

[25] J. Xiong, P. Zhang, T. You, C. Li, W. Huang, and Y. Zha, "Diffsal: Joint audio and video learning for diffusion saliency prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 27 273–27 283.

[26] Y. Ioannou, D. Robertson, R. Cipolla, and A. Criminisi, "Deep roots: Improving cnn efficiency with hierarchical filter groups," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[27] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[28] J. Pan, S. Chen, M. Z. Shou, Y. Liu, J. Shao, and H. Li, "Actor-context-actor relation network for spatio-temporal action localization," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 464–474.

[29] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *International Conference on Computer Vision (ICCV)*, 2019, pp. 6202–6211.

[30] C. Gu, C. Sun, D. A. Ross, C. Vondrick, C. Pantofaru, Y. Li, S. Vijayanarasimhan, G. Toderici, S. Ricco, R. Sukthankar, C. Schmid, and J. Malik, "Ava: A video dataset of spatio-temporally localized atomic visual actions," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6047–6056.

[31] S. Mathe and C. Sminchisescu, "Actions in the eye: Dynamic gaze datasets and learnt saliency models for visual recognition," *TPAMI*, vol. 37, no. 7, pp. 1408–1424, 2014.

[32] X. Min, G. Zhai, K. Gu, and X. Yang, "Fixation prediction through multimodal analysis," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 13, no. 1, 2016.

[33] 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.

[34] ——, "Toward the introduction of auditory information in dynamic visual attention models," in *2013 14th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, 2013, pp. 1–4.

[35] ——, "An efficient audiovisual saliency model to predict eye positions when looking at conversations," in *2015 23rd European Signal Processing Conference (EUSIPCO)*. IEEE, 2015, pp. 1531–1535.

[36] 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, pp. 5–24, 2011.

[37] P. Koutras, A. Katsamanis, and P. Maragos, "Predicting eyes' fixations in movie videos: Visual saliency experiments on a new eye-tracking database," in *Engineering Psychology and Cognitive Ergonomics*, D. Harris, Ed., 2014, pp. 183–194.

[38] F. Hu, S. Palazzo, F. P. Salanitri, G. Bellitto, M. Moradi, C. Spampinato, and K. McGuinness, "Tinyhd: Efficient video saliency prediction with heterogeneous decoders using hierarchical maps distillation," in *Winter Conference on Applications of Computer Vision (WACV)*, 2023.

[39] C. Ma, H. Sun, Y. Rao, J. Zhou, and J. Lu, "Video saliency forecasting transformer," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 10, pp. 6850–6862, 2022.

[40] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "What do different evaluation metrics tell us about saliency models?" vol. 41, no. 3, pp. 740–757, 2019.

[41] R. Agrawal, S. Jyoti, R. Girmaji, S. Sivaprasad, and V. Gandhi, "Does audio help in deep audio-visual saliency prediction models?" in *Proceedings of the 2022 International Conference on Multimodal Interaction*, 2022, pp. 48–56.