Lighter and Faster Two-Pathway CMRNet for Video Saliency Prediction

Sai Phani Kumar Malladi, Jayanta Mukhopadhyay, Mohamed-Chaker Larabi, Santanu Chaudhury

LIGHTER AND FASTER TWO-PATHWAY CMRNET FOR VIDEO SALIENCY PREDICTION

Sai Phani Kumar Malladi[⋆] *Jayanta Mukhopadhyay*[⋆] *Mohamed-Chaker Larabi*[†] *Santanu Chaudhury*[‡]

[⋆] Visual Information Processing Lab, Dept. of Computer Science & Engg., IIT Kharagpur, India
[†] XLIM UMR CNRS 7252, University of Poitiers, France
[‡] Dept. of Computer Science & Engg., IIT Jodhpur, India

ABSTRACT

Existing dynamic saliency prediction models face challenges like inefficient spatio-temporal feature integration, ineffective multi-scale feature extraction, and lacking domain adaptation because of huge pre-trained backbone networks. In this paper, we propose a two pathway architecture with effective feature integration of spatial and temporal domains at multiple scales for video saliency prediction. Frame and optical flow pathways extract features from video frame and optical flow maps, respectively using a series of cross-concatenated multi-scale residual (CMR) blocks. We name this network as two-pathway CMRNet (TP-CMRNet). Every CMR block follows a feature fusion and attention module for merging features from two pathways and guiding the network to weigh salient regions, respectively. A bi-directional LSTM module is used for learning the task by looking at previous and next video frames. We build a simple decoder for feature reconstruction into the final attention map. TP-CMRNet is comprehensively evaluated using three benchmark datasets: DHF1K, Hollywood-2, and UCF sports. We observe that our model performs at par with other deep dynamic models. In particular, we outperform all the other models with a lesser number of model parameters and lower inference time.

***Index Terms*—** Video saliency, two-pathway network, optical flow features, model parameters, inference time.

1. INTRODUCTION

Visual saliency prediction computes intra-frame saliency along with inter-frame motion and temporal information [1]. Unlike image analysis, video analysis has more challenges since motion and temporal information affect the attention of viewers [2]. Although the temporal domain brings rich motion information, complex motion patterns from the background, inconsistent movements among different foreground patterns, and camera motions make video saliency prediction more difficult [1]. Hence, it is important to effectively integrate features from multiple domains and scales which has been a long-time problem for dynamic saliency prediction. Traditional dynamic models [3, 4] are extensions of static models by just incorporating motion features. A multi-stream model with appearance, motion, and objectness together was proposed by Jiang et al. [5]. Recently, Wang et al. [6] used a CNN-LSTM network incorporated with supervised static attention mechanism for dynamic saliency prediction and achieved promising results. Two-stream network was first proposed in [7] for video action recognition, and in many spatio-temporal tasks [8]. Attention mechanisms allow the network focus on important aspects and have shown great successes in computer vision [9]. To summarize, we differ with previous works in two ways: (1) prior multi-stream models failed to extract multi-scale features within the operating blocks of the stream but generated features at different resolutions using max-pooling operations, and (2) prior models used huge pre-trained networks which suffer domain adaptation and did not extract features specific to video saliency.

In this work, we address the above issues with a deep network that handles inefficient feature integration of spatial and temporal domains at multiple scales. We propose to build a two-pathway network that extracts features from video frames and optical flow maps, and merge them using dense residual cross connections. In frame (FP) and optical flow (OFP) pathways, we employ an efficient feature extractor working at multiple scales by sharing information among them. We proposed this in [10] and use it here for effective global and local contextual information extraction. The proposed video saliency network is named as Two-Pathway CMRNet (TP-CMRNet) (Fig. 1) since it has cross-concatenated multi-scale residual (CMR) block as the building block. A feature fusion module follows every CMR block in FP that merges with corresponding features from OFP. Then, for a better learning, we propose to use an attention module to focus on salient regions. A bi-directional LSTM (BD-LSTM) [11] module is used which provides information back and forth about the sequence. As a whole, following are the major contributions from our work: (1) Two-pathway network for extracting multi-scale spatio-temporal features from video frames and optical flow information, (2) CMR block for efficient global and local contextual feature extraction, and (3) End-to-end trainable dynamic saliency model with less number of parameters and lower inference time than the existing models.

2. PROPOSED METHODOLOGY

Fig. 1 shows an overview of the proposed TP-CMRNet architecture based on CMR block [10]. We employ CMR blocks in both the pathways to extract multi-scale global and local contextual features with local residual learning. We also employ dense residual connections [12] between two pathways for comprehensive feature integration. To ensure that small movements are not discarded during feature encoding, we perform feature fusion [13]. Attention module helps the network learn to weigh the salient regions and aids in efficient task learning [14]. A BD-LSTM module allows the network learn both backward and forward information about the sequence [11]. We use a simple auto-encoder, before up-sampling the features, for efficient information transfer into the decoder. Then, we use a couple of convolution and sigmoid layers followed by up-sampling layer which together act as a decoder. We do not use any deconvolution layers in decoder since CMR block delivers denser features [15].

2.1. Frame and optical flow pathways

For a group of frames $\{V^t\}_{t=1}^{T}$, and the corresponding optical flow maps $\{F^t\}_{t=1}^{T}$, FP considers one input frame V^t and computes the frame features at different scales. However, OFP considers a group

of optical flow maps $\{F^t, F^{t+1}, ..., F^{t+S}\}$ from next S consecutive video frames and generate dynamic features as shown in Fig. 1.

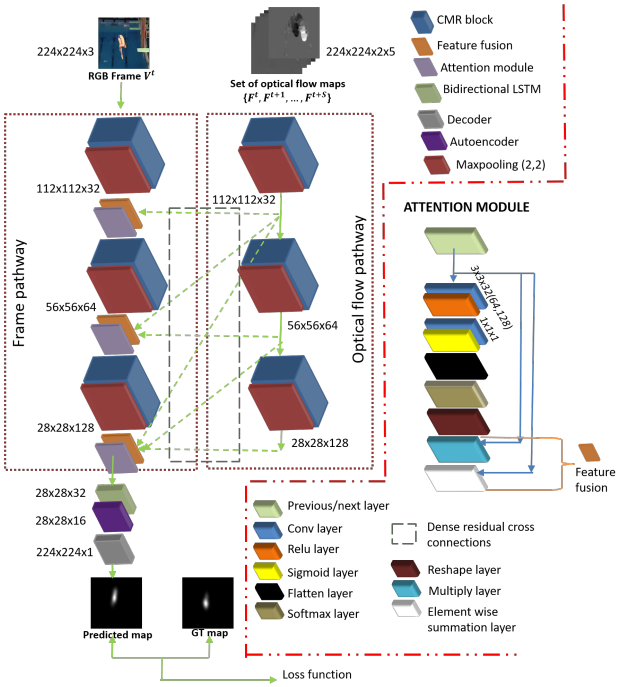

Fig. 1. Our proposed architecture and attention module.

2.1.1. *Cross-concatenated multi-scale residual (CMR) block*

CMR block [10] extracts efficient global and local contextual multi-scale features using inter-scale information sharing process. It concatenates features from multiple scales followed by a local residual learning process for better efficiency [16]. Two features at different scales, T_1 and F_1 undergo the above process and are passed again onto convolution layers at next level (Fig. 2). Then, the features T_2 and F_2 are processed and passed onto a $1 \times 1 \times 32$ convolution layer to reduce the number of parameters [16]. Now, the local residual learning computes element wise multiplication of the original feature with the one resulting after $1 \times 1 \times 32$ layer. Down-sampling operations were used in prior works to reduce computations but it results in low spatial resolutions. However, removing down-sampling operations results in reduced receptive field size. In order to trade-off between the computations and feature spatial resolution, we consider only two scales in CMR block [17].

We try to tightly incorporate the feature fusion among the two pathways using dense residual cross connections [12] between their corresponding features. Considering that FP tends to dominate OFP during training [2], we fuse the corresponding features F_f and F_o, respectively as:

$$F_f \leftarrow F_f + F_f \odot F_o \tag{1}$$

where $\odot$ represents the Hadamard product. This helps in potentially preserving the frame features though the corresponding optical flow features are infinitesimally small. It is similar to information exchange across various scales in our CMR block (Fig. 2).

2.1.2. *Attention module [18]*

The attention mechanism instructs the network to focus on certain salient features of the input at multiple scales. We adapt an attention module from [18] for learning the multi-scale fused features more powerfully. This module is embedded into FP hierarchically for enhancing multi-scale spatio-temporal salient features. Fig. 1 demonstrates the attention module learning an attention mask $A \in [0,1]^{W \times H}$ to softly weigh the spatio-temporal salient feature $STF \in R^{W \times H \times C}$ obtained after a feature fusion module in the frame pathway. A normalized mask is obtained by applying a softmax operation for highlighting the prominence of various regions in the feature. In the same way, we obtain attention masks at different scales throughout FP with highlighted multi-scale important regions. The feature fusion after *softmax* layer handles the cases when the attention module has not been well trained with selectiveness. Hence, the following identity mapping over feature STF after attention module is given as:

$$(STF)^c \leftarrow (STF)^c + A \odot (STF)^c \tag{2}$$

where $c = 1..C$ represents channels in STF. Eq. 2 also shows a residual structure for effective learning by avoiding new drastic changes into the original feature. This is also used as the feature fusion module in Fig. 1 for merging information from frame and optical flow pathways.

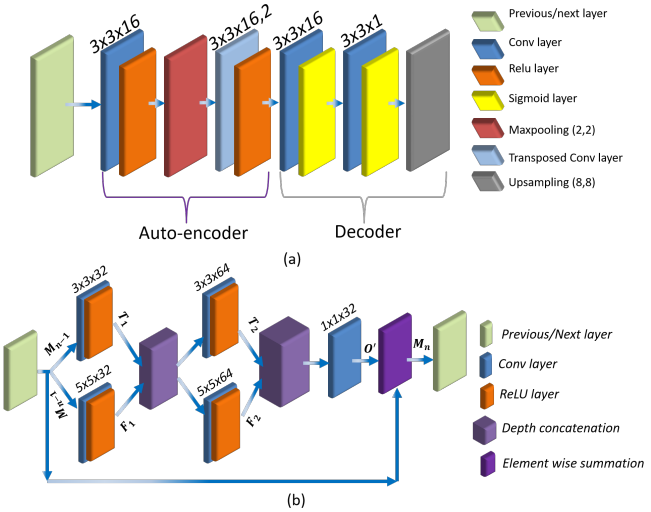

Fig. 2. (a) Auto-encoder & decoder, (b) CMR block [10].

2.1.3. *Bi-directional LSTM, auto-encoder, and decoder*

In order to temporally model the sequential data from video frames, we employ a bi-directional LSTM (BD-LSTM) [11]. It allows the network to have both the backward and forward information about the sequence at every time step. From Fig. 1, we see that feature shape is $28 \times 28 \times 32$ after the BD-LSTM module. Now, we propose to use a simple auto-encoder while we reduce the feature dimensions as shown in Fig. 2. Since, we aim to obtain a single channel saliency map at the end, abruptly reducing the number of feature channels results in information loss [19]. A simple auto-encoder helps in forming a smooth information flow from high dimensional features to low dimensional ones. In order to build the final saliency map, we use a simple decoder with a series of convolution and sigmoid activations. Then, we use a bi-cubic up-sampling layer to make the output resolution similar to the input for reduced computational cost.

2.1.4. *Loss function*

We build a loss function with four evaluation metrics measuring the quality of the computed saliency map ensuring the network learns saliency effectively. It is given as:

$$L(S,B,C) = L_{KL}(S,C) + \omega_1 L_{CC}(S,C) + \omega_2 L_{NSS}(S,B) + \omega_3 L_{SIM}(S,C) \quad (3)$$

where S, B, and C represent the predicted saliency map, ground truth (GT) binary fixation map, and continuous GT saliency map, respectively. L_{KL}, L_{CC}, L_{NSS}, and L_{SIM} represent the loss terms derived from KL divergence [20], linear correlation coefficient [21], normalized scanpath saliency (NSS) [21], and similarity (SIM) [20] metrics, respectively. Similar to [18], we choose ω_1, ω_2, and ω_3 as 0.2, 0.1, and 0.1, respectively.

3. EXPERIMENTAL EVALUATION

3.1. Datasets

In this work, we use the following three available benchmark video saliency datasets:

- **DHF1K [22]** includes 1000 videos with varied motion patterns and diverse contents of 7 main categories. The videos are split into 600, 100, and 300 for training, validation, and testing, respectively along with the gaze data of 17 observers.
- **Hollywood-2 [23]** includes 1707 videos performing 12 different activities. The videos are split into 823 for training, and 884 for testing along with the gaze data of 19 observers.
- **UCF Sports [24]** includes 150 videos performing 9 different sport actions with 103 videos for training, and 47 for testing.

3.2. Implementation details

We implement our network using Keras with Tensorflow back-end. Our network is trained from scratch and do not use any transfer learning. We use Adam optimizer [25] with an initial learning rate of 10^{-4} which is scaled down by a factor of 0.1 after every two epochs. We choose a video batch size of 1 and a frame batch size of 5 which means a group of 5 frames are taken from a single video as a training data batch. As a pre-processing step, we generate optical flow maps using Flownet 2.0 [26] before training. We consider the training data of all the three datasets as a whole for our training. We feed a group of consecutive optical flow maps to OFP when a video frame is fed to FP. We consider the following evaluation metrics for evaluating the model performance: AUC Judd (AUC-J) [21], shuffled AUC (s-AUC) [21], similarity (SIM), linear correlation coefficient (CC), and normalized scanpath saliency (NSS).

3.3. Model ablation analysis

Varied number of consecutive optical flow maps: Table 1 shows our ablation analysis results on the test sets of three benchmark datasets. Ablation results on UCF test-set ensures the robustness and effectiveness of our TP-CMRNet since sports videos contain abrupt motions and scene changes. Now, during this particular ablation study, our model is incorporated with feature fusion, attention, and BD-LSTM modules. By varying the number of optical flow maps, we find that there is a considerable performance improvement until 9 consecutive maps and it is not the case after that.

Feature fusion module: We evaluate our model performance by eliminating the feature fusion modules after every CMR block in FP. From the previous analysis, we decide to feed 9 consecutive flow maps into OFP for ablation analysis hereafter. We find a significant performance deterioration without feature fusion. This is because it tightly packs the spatial and temporal features from two pathways and helps even in detecting very small movements in videos.

Attention module: We study our model performance by eliminating the attention modules after every CMR block in FP, shown in Fig. 1. The performance is seen to be much better with attention module since it drives the network to learn weights corresponding to salient regions more prominently.

Replacing BD-LSTM with Conv. LSTM: We find that BD-LSTM efficiently learns to predict saliency by observing the information flow from both directions, as shown in Table. 1. At the end, we evaluate our model performance incorporating all the best possible results from the ablation studies and find its optimal performance.

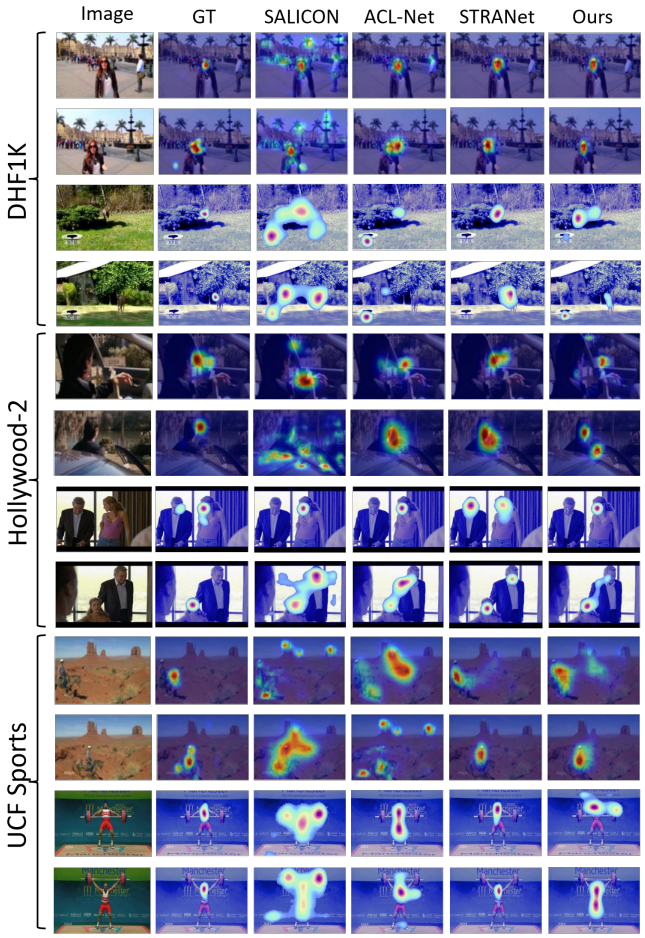

Fig. 3. Qualitative comparison with the SOTA models.

3.4. Comparison with state-of-the-art (SOTA) models

Comprehensive comparison: Table 3 shows that TP-CMRNet is lighter than every other model by at least 5.1 times. TASED-Net is found to be the next lighter model with 25 million parameters, while ours has 4.9 million. Our model takes an average inference time of 0.041 sec for predicting saliency map which is faster than other models by at least 1.8 times. This happens since our model does not use any part of the existing heavy deep architectures. In contrast, all the

Table 1. Model ablation analysis on test sets of UCF sports, DHF1K, and Hollywood-2 datasets.

Dataset →	UCF sports test set					DHF1K test set					Hollywood-2 test set				
Model ↓	AUC-J ↑	SIM ↑	s-AUC ↑	CC ↑	NSS ↑	AUC-J ↑	SIM ↑	s-AUC ↑	CC ↑	NSS ↑	AUC-J ↑	SIM ↑	s-AUC ↑	CC ↑	NSS ↑
Impact of choosing varied number of consecutive optical flow maps															
3	0.861	0.474	0.727	0.557	2.920	0.862	0.318	0.666	0.423	2.462	0.878	0.505	0.740	0.621	3.248
5	0.872	0.488	0.738	0.568	2.941	0.873	0.332	0.677	0.434	2.483	0.889	0.519	0.751	0.632	3.269
7	0.883	0.499	0.749	0.579	2.962	0.884	0.343	0.686	0.445	2.504	0.900	0.530	0.764	0.643	3.290
9	**0.892**	**0.503**	**0.755**	**0.586**	**2.970**	**0.892**	**0.347**	**0.692**	**0.452**	**2.512**	**0.908**	**0.533**	**0.770**	**0.650**	**3.298**
11	0.894	0.503	0.758	0.580	2.968	0.892	0.349	0.695	0.448	2.510	0.907	0.531	0.767	0.653	3.295
Impact of feature fusion module															
w/o feature fusion	0.878	0.493	0.738	0.568	2.948	0.843	0.311	0.625	0.410	2.440	0.859	0.497	0.703	0.608	3.226
Impact of attention module															
w/o attention module	0.860	0.480	0.713	0.556	2.923	0.861	0.324	0.650	0.422	2.465	0.877	0.510	0.728	0.620	3.251
Impact of Conv. LSTM (in place of Bi-directional Conv. LSTM)															
with Conv. LSTM	0.873	0.484	0.721	0.560	2.930	0.874	0.328	0.658	0.426	2.472	0.89	0.514	0.736	0.624	3.258
Our Model															
9 flow maps + feature fusion + attention module + BD-LSTM	**0.892**	**0.503**	**0.755**	**0.586**	**2.970**	**0.892**	**0.347**	**0.692**	**0.452**	**2.512**	**0.908**	**0.533**	**0.770**	**0.650**	**3.298**

Table 2. Quantitative comparison of our model with the SOTA deep video saliency models.

Dataset →	DHF1K					Hollywood-2					UCF Sports				
Method ↓	AUC-J ↑	SIM ↑	s-AUC↑	CC↑	NSS	AUC-J↑	SIM↑	s-AUC↑	CC↑	NSS↑	AUC-J↑	SIM↑	s-AUC↑	CC↑	NSS↑
SALICON [27]	0.857	0.232	0.590	0.327	1.901	0.856	0.321	0.711	0.425	2.013	0.848	0.304	0.738	0.375	1.838
OM-CNN [28]	0.856	0.256	0.583	0.344	1.911	0.887	0.356	0.693	0.446	2.313	0.870	0.321	0.691	0.405	2.089
ACL-Net [6]	0.890	0.315	0.601	0.434	2.354	0.913	**0.542**	0.757	0.623	3.086	*0.905*	0.496	*0.767*	*0.603*	*3.200*
TASED-Net [29]	**0.895**	**0.361**	**0.712**	**0.470**	*2.558*	*0.918*	0.507	0.768	0.646	*3.302*	0.899	0.469	0.752	0.582	2.920
STRA-Net [18]	**0.895**	*0.355*	0.663	*0.458*	**2.667**	**0.923**	*0.536*	**0.774**	**0.662**	**3.478**	**0.914**	**0.535**	**0.790**	**0.645**	**3.472**
TP-CMRNet (Ours)	*0.892*	0.347	*0.692*	0.452	2.512	0.908	0.533	*0.770*	*0.650*	3.298	0.892	*0.503*	0.755	0.586	2.970

other models in Table 3 use either a portion or the whole architecture from networks like AlexNet, ResNet50, and GoogleNet. Hence, our model outperforms all the other deep visual saliency models in terms of parameters and inference time.

Table 3. Comprehensive comparison with the SOTA deep dynamic saliency models

Model	# parameters ($\times 10^6$)	Avg. inference time (per video frame, in sec)
SALICON [27]	∼206	0.88
OM-CNN [28]	∼80	0.13
ACL-Net [6]	∼140	0.19
TASED-Net [29]	∼25	0.21
STRA Net [18]	∼46	0.074
TP-CMRNet (Ours)	**4.9**	**0.041**

Quantitative comparison: The quantitative evaluation of TP-CMRNet (Table 2) highlights the best and second best values of a metric in bold and underlined forms, respectively. We find that our model performs at par with others but may not outperform. On average, our model varies by 3%, 3.8%, 2.8%, 3.8%, and 4% when compared with the best. This margin is considerably smaller since our motive is to make it lighter and faster.

Qualitative comparison: Fig. 3 shows a few qualitative results from three data sets where we provide two video frames of a clip for demonstration. Though there are varying scales and backgrounds in DHF1K videos, our network predicts the saliency at par with SOTA models. We attribute this to the efficient multi-scale feature extracting CMR block and dense residual connection. From Hollywood-2 results, our model even keeps a track of smaller scene objects just like STRA-Net, whereas others track both moving and salient objects. This happens since the attention mask weighs features distinguishing salient regions more prominently. Even for UCF sports videos, we observe a comparably good performance though they contain abrupt motion and diverse content. As a whole, on three benchmark datasets, we witness a promising performance of our lighter and faster TP-CMRNet which stands alone without any huge pre-trained networks.

4. CONCLUSION

In this work, we propose TP-CMRNet, a two-pathway deep model that efficiently integrates spatial and temporal domain features with effective multi-scale feature extracting CMR block using dense residual cross connections. Features from frame and optical flow pathways are tightly integrated using feature fusion and the attention module guides the network towards learning the salient region information. A BD-LSTM module after two pathways helps TP-CMRNet have both the forward and backward sequence information for better performance. A simple auto-encoder is used, to avoid abrupt reduction in feature dimensions and smooth information flow. A decoder with a series of convolution with sigmoid activation and a bi-cubic up-sampling layer reconstructs the final prediction map. TP-CMRNet is trained on training sets of DHF1K, Hollywood-2, and UCF sports datasets and evaluated on their test sets. TP-CMRNet performs comparatively better than a few SOTA models but may not outperform the best. However, TP-CMRNet excels in terms of parameters and inference time by outperforming all the relevant deep dynamic models.

5. ACKNOWLEDGEMENT

This work is carried out under the Institute Challenge Grants, IIT Kharagpur sponsored project entitled as "Visual attention assisted image and video compression" (Grant No. IIT/SRIC/CS/VIV_ICG_2017_SGCIR/2018-19/085).

6. REFERENCES

[1] Pingping Zhang, Wei Liu, Dong Wang, Yinjie Lei, Hongyu Wang, and Huchuan Lu, "Non-rigid object tracking via deep multi-scale spatial-temporal discriminative saliency maps," *Pattern Recognition*, vol. 100, pp. 107130, 2020.

[2] Xiaoqiang Lu, Yaxiong Chen, and Xuelong Li, "Hierarchical recurrent neural hashing for image retrieval with hierarchical convolutional features," *IEEE transactions on image processing*, vol. 27, no. 1, pp. 106–120, 2017.

[3] Laurent Itti, Nitin Dhavale, and Frederic Pighin, "Realistic avatar eye and head animation using a neurobiological model of visual attention," in *Applications and Science of Neural Networks, Fuzzy Systems, and Evolutionary Computation VI*. SPIE, 2003, vol. 5200, pp. 64–78.

[4] Lingyun Zhang, Matthew H Tong, and Garrison W Cottrell, "Sunday: Saliency using natural statistics for dynamic analysis of scenes," in *Proceedings of the 31st annual cognitive science conference*. AAAI Press Cambridge, MA, 2009, pp. 2944–2949.

[5] Lai Jiang, Mai Xu, Tie Liu, Minglang Qiao, and Zulin Wang, "Deepvs: A deep learning based video saliency prediction approach," in *Proceedings of the european conference on computer vision (eccv)*, 2018, pp. 602–617.

[6] Wenguan Wang, Jianbing Shen, Fang Guo, Ming-Ming Cheng, and Ali Borji, "Revisiting video saliency: A large-scale benchmark and a new model," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4894–4903.

[7] Karen Simonyan and Andrew Zisserman, "Two-stream convolutional networks for action recognition in videos," *Advances in neural information processing systems*, vol. 27, 2014.

[8] R Christoph and Feichtenhofer Axel Pinz, "Spatiotemporal residual networks for video action recognition," *Advances in neural information processing systems*, pp. 3468–3476, 2016.

[9] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang, "Residual attention network for image classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3156–3164.

[10] Sai Phani Kumar Malladi, Jayanta Mukhopadhyay, Chaker Larabi, and Santanu Chaudhury, "Lighter and faster cross-concatenated multi-scale residual block based network for visual saliency prediction," in *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021, pp. 2503–2507.

[11] Zhiheng Huang, Wei Xu, and Kai Yu, "Bidirectional lstm-crf models for sequence tagging," *arXiv preprint arXiv:1508.01991*, 2015.

[12] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu, "Residual dense network for image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2472–2481.

[13] Christoph Feichtenhofer, Axel Pinz, and Richard P Wildes, "Spatiotemporal multiplier networks for video action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4768–4777.

[14] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Haibin Lin, Zhi Zhang, Yue Sun, Tong He, Jonas Mueller, R Manmatha, et al., "Resnest: Split-attention networks," *arXiv preprint arXiv:2004.08955*, 2020.

[15] Augustus Odena, Vincent Dumoulin, and Chris Olah, "Deconvolution and checkerboard artifacts," *Distill*, vol. 1, no. 10, pp. e3, 2016.

[16] Sheng Yang and Lin, "A dilated inception network for visual saliency prediction," *IEEE Transactions on Multimedia*, 2019.

[17] Samuel F Dodge and Lina J Karam, "Visual saliency prediction using a mixture of deep neural networks," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 4080–4090, 2018.

[18] Qiuxia Lai, Wenguan Wang, Hanqiu Sun, and Jianbing Shen, "Video saliency prediction using spatiotemporal residual attentive networks," *IEEE Transactions on Image Processing*, vol. 29, pp. 1113–1126, 2019.

[19] Changqing Zhang, Yeqing Liu, and Huazhu Fu, "Ae2-nets: Autoencoder in autoencoder networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 2577–2585.

[20] Tilke Judd, Frédo Durand, and Antonio Torralba, "A benchmark of computational models of saliency to predict human fixations," 2012.

[21] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara, "Predicting human eye fixations via an lstm-based saliency attentive model," *IEEE TIP*, vol. 27, no. 10, pp. 5142–5154, 2018.

[22] Wenguan Wang, Jianbing Shen, Jianwen Xie, Ming-Ming Cheng, Haibin Ling, and Ali 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.

[23] Heng Wang and Cordelia Schmid, "Action recognition with improved trajectories," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 3551–3558.

[24] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.

[25] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[26] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox, "Flownet 2.0: Evolution of optical flow estimation with deep networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2462–2470.

[27] Xun Huang, Chengyao Shen, Xavier Boix, and Qi Zhao, "Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 262–270.

[28] Lai Jiang, Mai Xu, and Zulin Wang, "Predicting video saliency with object-to-motion cnn and two-layer convolutional lstm," *arXiv preprint arXiv:1709.06316*, 2017.

[29] Kyle Min and Jason J Corso, "Tased-net: Temporally-aggregating spatial encoder-decoder network for video saliency detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2394–2403.