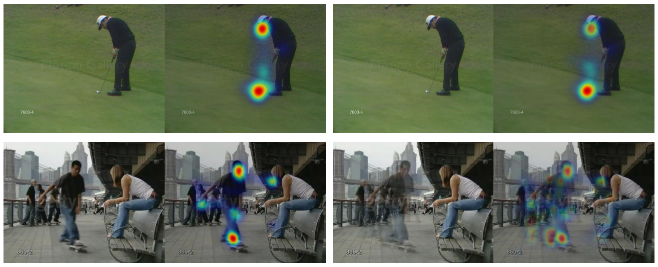
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2021.3094974, IEEE Transactions on Cognitive and Developmental Systems

# A Gated Fusion Network for Dynamic Saliency Prediction

Aysun Kocak, Erkut Erdem and Aykut Erdem

***Abstract*—Predicting saliency in videos is a challenging problem due to complex modeling of interactions between spatial and temporal information, especially when ever-changing, dynamic nature of videos is considered. Recently, researchers have proposed large-scale datasets and models that take advantage of deep learning as a way to understand what's important for video saliency. These approaches, however, learn to combine spatial and temporal features in a static manner and do not adapt themselves much to the changes in the video content. In this paper, we introduce Gated Fusion Network for dynamic saliency (GFSalNet), the first deep saliency model capable of making predictions in a dynamic way via gated fusion mechanism. Moreover, our model also exploits spatial and channel-wise attention within a multi-scale architecture that further allows for highly accurate predictions. We evaluate the proposed approach on a number of datasets, and our experimental analysis demonstrates that it outperforms or is highly competitive with the state of the art. Importantly, we show that it has a good generalization ability, and moreover, exploits temporal information more effectively via its adaptive fusion scheme.**

***Index Terms*—dynamic saliency estimation, gated fusion, deep saliency networks**



A single input frame and its corresponding fixation map

Four consecutive overlaid frames and their overlaid fixation maps

Fig. 1: Predicting video saliency requires finding a harmonious interaction between appearance and temporal information. For example, while the first row shows a case in which attention is guided more by visual appearance, in the second row, motion is the most determining factor for attention. Hence, we speculate that an adaptive scheme would be better suited for this task.

## I. Introduction

Human visual system employs visual attention mechanisms to effectively deal with huge amount of information by focusing only on salient or attention grabbing parts of a scene, and thus filtering out irrelevant stimuli. Saliency estimation methods offer different computational models of attention to mimic this key component of our visual system. These methods generate a so-called saliency map within which a pixel value indicates the likelihood of that pixel being fixated by a human. Since the pioneering work of [1], this research area has gained a lot of interest in the last few decades (please refer to [2] for an overview), and it has found to have practical use in a variety of computer vision tasks such as visual quality assessment [3], [4], image and video resizing [5], [6], video summarization [7], to name a few. Early saliency prediction approaches use low-level (color, orientation, intensity) and/or high-level (pedestrians, faces, text, etc.) image features to estimate salient regions. While low-level cues are used to detect regions that are different from their surroundings, top-down cues are used to infer high-level semantics to guide the model. For example, humans tend to focus some object classes more than others. Recently, deep learning based models have started to dominate over the traditional approaches as they can directly learn both low and high-level features relevant for saliency prediction [8], [9].

Most of the literature on saliency estimation focuses on static images. Lately, predicting saliency in videos has also gained some attraction, but it still remains a largely unexplored field of research. Video saliency models (also called dynamic saliency models) aim to predict attention grabbing regions in dynamically changing scenes. While static saliency estimation considers only low-level and high-level spatial cues, dynamic saliency needs to take into account temporal information too as there is evidence that moving objects or object parts can also guide our attention. Motion and appearance play complementary roles in human attention and their significance can change over time. As we illustrate in Fig. 1, in dynamic scenes, humans tend to focus more on moving parts of the scene and the eye fixations change over time, showing the importance of motion cues (bottom row). On the other hand, when there is practically no motion in the scene, low-level appearance cues dominantly guide our attention and we focus more on the regions showing different visual characteristics than their surroundings (top row). Motivated by these observations, in this work, we develop a deep dynamic saliency model which handles spatial and temporal changes in the visual stimuli in an adaptive manner.

The first generation of dynamic saliency methods were simply extensions of the static saliency approaches, *e.g.* [10], [11], [12], [13], [14]. In other words, these methods adapted the strategies proposed for static scenes and mostly modified them to work on either 3D feature maps that are formed by stacking 2D spatial features over time or 2D feature maps encoding motion information like optical flow images. Several follow-up works, however, have approached the problem from a fresh perspective and developed specialized methods for dynamic saliency detection, *e.g.* [15], [16], [17], [18], [19], [20], [21], [22], [23]. These models either utilize novel spatio-temporal features or employ data-driven techniques to learn relevant features from data. As with the case of state-of-the-art static saliency models, approaches based on deep learning have also shown promise for dynamic saliency. These studies basically explore different neural architectures used for processing temporal and spatial information in a joint manner, and they either use 3D convolutions [24], LSTMs [24], [25] or multi-stream architectures that encode temporal information separately [26], [27], [28].

In this work, we introduce Gated Fusion Network for video saliency (GFSalNet). Our proposed network model is radically different from the previously proposed deep models in that it includes a novel content-driven fusion scheme to combine spatial and temporal streams in a more dynamic manner. In particular, our model is based on two-stream CNNs [29], [30], which have been successfully applied to various video analysis tasks. To our interest, these architectures are inspired by the ventral and dorsal pathways which are suggested to subserve object identification and motion perception, respectively [31], [32], in the human visual cortex [33]. Although the use of two-stream CNNs in video saliency prediction has been investigated before [27], the main novelty of our work lies in the ability to fuse appearance and motion information in a spatio-temporally coordinated manner by estimating the importance of each cue with respect based on the current video content.

The rest of the paper is organized as follows: In Section 2, we give a brief overview of the existing dynamic saliency approaches. In Section 3, we present the details of our proposed deep architecture for video saliency. In Section 4, we give the details of our experimental setup, including evaluation metrics, datasets and the competing dynamic saliency models, and discuss the results of our experiments. Finally, in the last section, we offer some concluding remarks.

Our codes and predefined models, along with the saliency maps extracted with our approach, will be publicly available at the project website[1].

## II. Related Work

Early visual saliency models can be dated back to 1980s with the Feature Integration Theory by [34]. The first models of saliency, such as [35], [1], provide computational solutions to [34], and since then a notable number of saliency models are developed, most of which deal with static scenes. For a detailed list of pre-deep learning saliency estimation approaches, please refer to [2]. After the availability of large-scale datasets, researchers proposed various deep learning based models for static saliency that outperformed previous approaches by a large margin [36], [37], [38], [39], [40], [41], [42], [43], [44].

**Early models for dynamic saliency** generally depend on previously proposed static saliency models. Adaptation of these models to dynamic scenes is achieved by considering features related to motion such as the optical flow information. For example, [10] proposed a saliency prediction method called PQFT that predicts the salient regions via the phase spectrum of Fourier Transform of the given image. In particular, PQFT generates a quaternion image representation by using color, intensity, orientation and motion features and estimates the salient regions in the frequency domain by using this combined representation. [11] extracted salient parts of video frames by similarly performing a spectral analysis of the frames considering both spatial and temporal domains. [12] employed local regression kernels as features to calculate self similarities between pixels or voxels for figure-ground segregation. [13] extended the previously proposed static saliency model by [45]'s model by including motion cues to the graph-theoretic formulation. [46] employ a two stream approach that generates spatial saliency map (using color and texture features) and temporal saliency map (using optical flow feature) separately and combines these maps with an entropy based adaptive method. [14] proposed a dynamic saliency model for activity recognition that works in an unsupervised manner. Their method is based on an encoding scheme that considers color along with motion cues.

Following these early approaches, the researchers started to develop novel video saliency models specifically designed for dynamic stimuli. For instance, [15] proposed a sparsity based framework that generates spatial saliency maps and temporal saliency maps separatelty based on entropy gain and temporal consistency, respectively, and then combines them. [16] integrated several visual cues such as static and dynamic image features based on color, texture, edge distribution, motion boundary histograms, through learning-based fusion strategies and later employed this dynamic saliency model for action recognition. [17] suggested a learning-based model that generates a candidate set regions with the use of existing methods and then predicts gaze transitions over subsequent video frames conditionally on these regions. [18] proposed a simple dynamic saliency model that combines spatial saliency maps with temporal saliency using pixel-wise maximum operation. In their work, while the spatial saliency maps are extracted using multi-scale analysis of low-level features, temporal saliency maps are obtained by examining dynamic consistency of motion through an optical flow model. [19] suggested an approach that independently estimates superpixel-level and pixel-level temporal and spatial saliency maps and subsequently combines them using an adaptive fusion strategy. [20] proposed an approach that oversegments video frames by using both spatial and temporal information and estimates the saliency score for each region by computing the regional contrast values via low-level features extracted from these regions. [21] suggested to learn a filter bank from low-level features for fixations. This filterbank encodes the association between local feature patterns and probabilities of human fixations, and is used to re-weight fixation candidates. [22] formulated another dynamic saliency model by exploiting the compressibility principle. More recently, [23] proposed a saliency model (called AWS-D) for dynamic scenes by considering the observation that high-order statistical structures carry most of the perceptually

[1] https://hucvl.github.io/GFSalNet/

relevant information. AWS-D [23] removes the second-order information from input sequence via a whitening process. Then, it computes bottom-up spatial saliency maps using a filter bank at multiple scales, and temporal saliency maps with the use of a 3D filter bank. Finally, it combines all these maps by considering their relative significance.

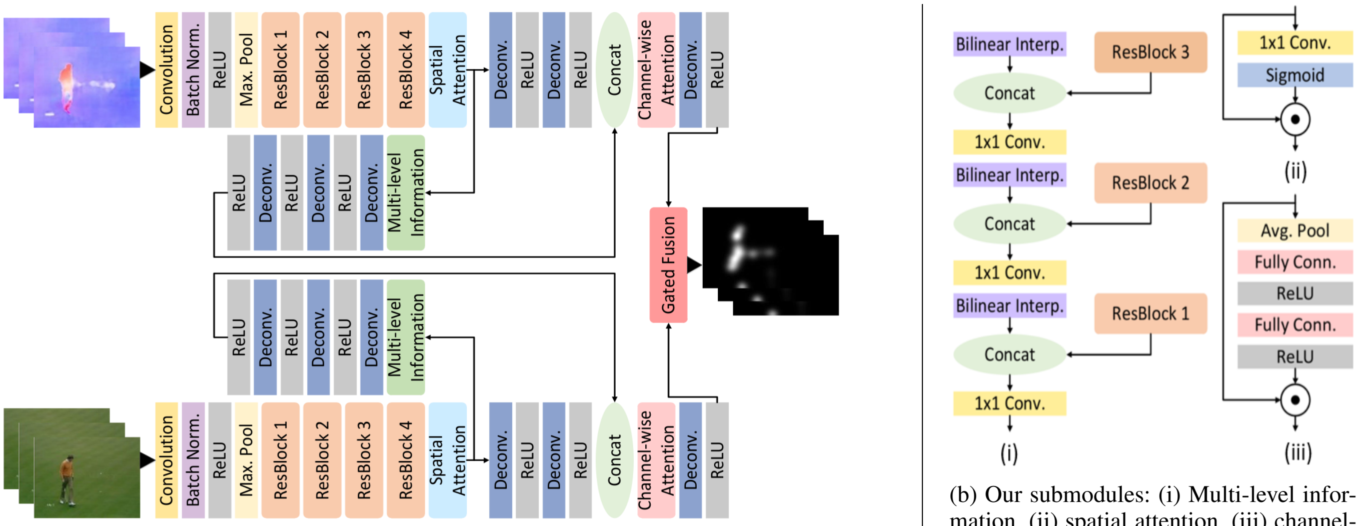
In addition to the aforementioned studies, some researchers also investigated the problem of salient object detection in videos where the main aim is not to predict human fixation maps in each frame but to detect foreground objects and their boundaries that pop out as compared to their surroundings [47], [48], [49], [50],  [51], [52], [53]. Some of the deep salient object detection methods also uses global and local information by processing information at multiple levels [54], [55], [56], [57], [58], [59], [60], [61]. Since, these methods are trained on salient object segmentation datasets and evaluated differently than the saliency prediction models, we do not include these studies in our experimental evaluation.

**Deep learning based dynamic saliency models** have received attention only recently. [24] proposed a recurrent mixture density network (RMDN) for spatio-temporal visual attention. The method uses a C3D architecture [62] as a backbone to integrate spatial and temporal information. This representation module is fed to a Long Short-Term Memory (LSTM) network, which is connected to Mixture Density Network (MDN) whose outputs are the parameters of a Gaussian mixture model expressing the saliency map of each frame. [27] suggested a two stream CNN model [29], [30] which considers the motion and appearance clues in videos. While, optical flow images are used to feed the temporal stream, raw RGB frames are used as input for the spatial stream. [26] presented an attention network to predict where driver is focused. In this work, the authors also proposed a dataset that consists of ego-centric and car-centric driving videos and eye tracking data belongs to the videos. Their network consists of three independent paths, namely spatial, temporal and semantic paths. While the spatial path uses raw RGB data as input, the temporal one uses optical flow data to integrate motion information and the last one processes the segmentation prediction on the scene given by the model by [63]. In the final layer of the network, the three independent maps are summed and then normalized to obtain the final saliency map. [28] proposed a deep model called OM-CNN which consists of two subnetworks, namely objectness subnet to highlight the regions that contain an object, motion subnet to encode temporal information, whose outputs are then combined to generate some spatio-temporal features. [25] proposed a model called ACLNet which employs a CNN-LSTM architecture to predict human gaze in dynamic scenes. The proposed approach focuses static information with an attention module and allows an LSTM to focus on learning dynamic information. Recently, [64] proposed an encoder-decoder based deep neural network called SalEMA, which employs a convolutional recurrent neural network method to include temporal information. In particular, it processes a sequence of RGB video frames as input to employ spatial and temporal information with the temporal information being inferred by the weighted average of the convolution state of the current frame and all the previous frames. [65] suggested a different model called TASED-Net, which utilizes a 3D fully-convolutional encoder-decoder network architecture where the encoded features are spatially upsampled while aggregating the temporal information. [66] recently developed another two-stream spatiotemporal salieny model called STRA-Net that considers dense residual cross connections and a composite attention module.

The aforementioned dynamic saliency models suffer from different drawbacks. The early methods employ (hand-crafted) low-level features that do not provide a high-level understanding of the video frames. Deep models eliminate this pitfall by utilizing an end-to-end learning strategy and, hence, provide better saliency predictions. They differ from each other by how they include motion information within their respective architectures. As we reviewed, the two main alternative approaches include using recurrent connections or processing data in multiple streams. Although RNN-based models help to encode temporal information with less amount of parameters, the encoding procedure compresses all the relevant information into a single vector representation, which affects the robustness especially for longer sequences. In that respect, the accuracy of the two-stream models do not, in general, degrade as the length of a sequence increases. Moreover, they are more interpretable as they need to perform fusion of spatial and temporal features in an explicit manner. On the other hand, their performance depends on accurate estimation of the optical flow maps used as input to the temporal stream. Hence, most of these two-stream models employ recent deep-learning based optical flow estimation models and even some of them uses some additional post-processing steps such as confining the absolute values of the magnitudes within a certain interval to avoid noise, as in STRA-Net [66]. Our proposed model also uses a two-stream approach, but as we will show, it exploits a novel and more dynamic fusion strategy, which boosts the performance and further improves the interpretability.

## III. Our Model

A general overview of our proposed spatio-temporal network architecture is given in Fig. 2(a). We use a two-stream architecture that processes temporal and spatial information in separate streams, similar to the one in [27]. That is, we respectively feed the spatial stream and temporal stream with RGB video frames and the corresponding optical flow images as inputs. Different than [27], however, our network combines information coming from several levels (Section III-A) and fuses both streams via a novel dynamic fusion strategy (Section III-C). We additionally utilize attention blocks (Section III-B) to select more relevant features to further boost the performance of our model. Here, we use a pre-trained ResNet-50 model [67] as the backbone of our saliency network as commonly explored by the previous saliency studies. In particular, we remove the average pooling and fully connected layers after the last residual block (ResBlock4) and then adapt it for saliency prediction by adding extra blocks. Using ResNet-50 model allows us to encode both low-, mid- and high-level cues in the visual stimuli in an efficient manner. Moreover, the number of network parameters is much smaller as compared to other alternative backbone networks.



(a) Our full model

(b) Our submodules: (i) Multi-level information, (ii) spatial attention, (iii) channel-wise attention blocks.

Fig. 2: Our two-stream dynamic saliency model uses RGB frames for spatial stream and optical flow images for temporal stream. These streams are integrated with a dynamic fusion strategy that we referred to as gated fusion. Our architecture also employs multi-level information block to fuse multi-scale features extracted at different levels of the network and attention blocks for feature selection. While the spatial attention block defines spatial importance weights for individual feature maps, the channel-wise attention block introduces feature-level weighting which allows for a better use of context information.

### A. *Multi-level Information Block*

As its name implies, the purpose of multi-level information block is to let the information extracted at different levels guide the saliency prediction process. It has proven to be useful that employing a multi-level/multiscale structure almost always improves the performance for many different vision tasks such as object detection [68], segmentation [69], [70], [71], and static saliency detection [72], [73]. In our work, we also employ a multi-level information block to enhance feature learning capability of our model. Specifically, it allows low-, mid-, and high-level information to be fused together and to be taken into account simultaneously while making predictions.

Fig. 2b-(i) shows the proposed multi-level information block that we employ in our model. This block considers low-level and high-level representations of frames by processing features maps which are extracted at each residual block. The aim is to combine primitive image features (*e.g.* edges, shared common patterns) obtained at lower levels with rich semantic information (*e.g.* object parts, faces, text) extracted at higher levels of the network. Here, we prefer to utilize $1 \times 1$ convolution and bilinear interpolation layers to combine cues from higher and lower levels. That is, after each residual block, we expand the feature map with bilinear interpolation to make equal size of the feature map with the size of the output of the previous residual block. Then, we concatenate the expanded feature map with the previous residual block's output and fuse them via $1 \times 1$ convolution layers.

### B. *Attention Blocks*

Neural attention mechanisms allow for learning to pay attention to features more useful for a given task, and hence, it has been demonstrated many times that they can boost the performance of a neural network architecture proposed for any computer vision problem, such as object detection [74]), visual question answering [75], pose estimation [76], image captioning [77] and salient object detection [72]. Motivated with these observations, in our work, we integrate several attention blocks to our proposed deep architecture to let the model choose the most relevant features for the dynamic saliency estimation problem. Resembling the structures in [77], [72], we exploit two separate attention mechanisms: *spatial* and *channel-wise* attention, as explained below.

Fig. 2b-(ii) shows our spatial attention block, which we introduce at the lower levels of our network model (see Fig. 2a) that helps to filter out the irrelevant information. The block takes the output of ResBlock4, shaped $[B \times C \times H \times W]$ with $C = 2048$, as input and it determines the important locations by calculating a weight tensor, which is shaped $[B \times 1 \times H \times W]$. To estimate this tensor, input channels are fused via $1 \times 1$ convolution layer following by a sigmoid layer. The output (shaped $[B \times C \times H \times W]$) of this block is a result of Hadamard product between input and spatial weight tensor.

The second type of our attention block, the channel-wise attention block, is shown in Fig. 2b-(iii), whose main purpose is to utilize the context information in a more efficient way. The block consists of average pooling, full connected and ReLU layers. In particular, it takes the concatenation of the feature maps from the main stream and multi-level information block as input which is shaped $[B \times 96 \times H \times W]$, then downsamples it with average pooling (output shape is $[B \times 96]$). The weight of each channel is determined after two fully connected layers followed by ReLUs. The shape of the matrices are $[B \times 24]$ and $[B \times 96]$ respectively. The output of last ReLU which is shaped $[B \times 96 \times 1 \times 1]$, contains a scalar value to weight

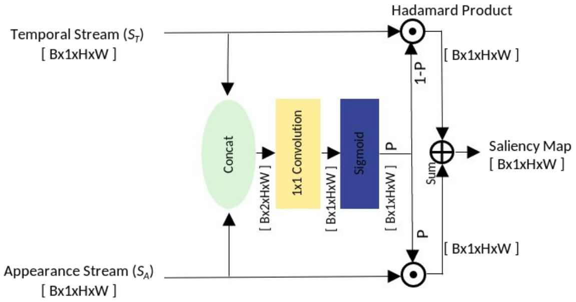


Fig. 3: Gated fusion block. It integrates the spatial and temporal streams to learn a weighted gating scheme to determine their contributions in predicting dynamic saliency of the current input video frame.

each channel. At the end of the block, the input feature map is weighted via Hadamard product.

### C. Gated Fusion Block

One of the main contributions of our framework is to employ a dynamic fusion strategy to combine temporal and spatial information. Gated fusion has been exploited before for different problems such as image dehazing [78], image deblurring [79], semantic segmentation [80]. The main purpose to use a gated fusion block is to combine different kind of information with a dynamic structure which considers the current inputs' characteristics. For example, in [80] feature maps that are generated via RGB information and depth information is combined for solving semantic segmentation. In our case, our aim is to come up with a fusion module that considers the content of the video at inference time. To our knowledge, we are the first to provide a truly dynamic approach for dynamic saliency. As opposed to the classical learning based approaches that learn the contributions of temporal and spatial streams in a static manner from the training data, our gated fusion block performs the fusion process in an adaptive way. That is, it decides the contribution of each stream on a location- and time-aware manner according to the content of the video.

The structure of the proposed gated fusion block is shown in Fig. 3. It takes the feature maps of the spatial and temporal streams as inputs and produces a probability map which is used to designate contribution of each stream with regard to their current characteristics. Let $S_A$, $S_T$ denote the feature maps from spatial and temporal streams, respectively. Gated fusion module first concatenates these features and then learns their correlations by applying a $1 \times 1$ convolution layer. After that, it uses a sigmoid layer to regularize the feature map which is used to estimate weights of the gate. Let $G_A$ and $G_T$ denote how confidently we can rely on appearance and motion, respectively, as follows:

$$G_A = P \ , \quad G_T = 1 - P \ , \tag{1}$$

where $P$ is the output of the sigmoid layer. Then, gated fusion module estimates the weights denoting the contributions of the spatial and temporal streams, as given below:

$$S'_A = S_A \odot G_A \ , \quad S'_T = S_T \odot G_T \ , \tag{2}$$

where $\odot$ represents the Hadamard product operation. Finally, it generates the final saliency map, $S_{final}$, via weighting the appearance and temporal streams' feature maps with the estimated probability map:

$$S_{final} = S'_A + S'_T \ . \tag{3}$$

As mentioned earlier, appearance and motion are the two important cues affecting attended regions in videos. Fig. 4 visualizes how gated fusion block adaptively integrates these two visual modalities on two sample video sequences. While the appearance stream computes a saliency map $S_A$ from the RGB frame, the temporal stream extracts a second saliency map $S_T$ from the optical image obtained from successive frames. As can be seen, these intermediate maps encode different characteristic of the input dynamic stimuli. The appearance based saliency map $S_A$ mostly focuses on the regions that have distinct visual properties than theirs surroundings, whereas the motion based saliency map $S_T$ mainly pay attention to motion. Gated fusion scheme estimates spatially varying probability maps $G_A$ and $G_T$ and employs them to integrate the appearance and temporal streams, respectively, resulting in more confident predictions. The spatial stream generally gives more accurate predictions than the temporal stream, as will be presented in the Experiments section. On the other hand, as can be seen from the estimated weight maps $G_A$ and $G_T$, the gated fusion scheme in the proposed model has a tendency to pay more attention to the temporal stream. We suspect that this is because the model considers that it may carry auxiliary information. In that regard, it can be also argued that the proposed gated fusion block improves the interpretability of our deep model on a given visual stimuli via the estimated probabilty maps as they allow us to highlight which regions are ignored or paid more attention by the appearance and the temporal streams throughout the sequence.

## IV. Experiments

Here, we first provide a brief review of the datasets used in our experimental analysis. Then, we give the details of our training procedure including the loss functions and settings we use to train our proposed model. Next, we summarize the evaluation metrics and the dynamic saliency models used in our experiments. We then discuss our findings and present some qualitative and quantitative results. Finally, we present an ablation study to evaluate the effectiveness of the blocks of the proposed dynamic saliency model.

### A. Datasets

In our experiments, we employ six different datasets to evaluate the effectiveness of the proposed saliency model. The first four, namely UCF-Sports [81], Holywood-2 [82], DHF1K [25], and DIEM [83], are the most commonly used benchmarks. Among them, we specifically utilize DIEM to test the generalization ability of our model. The last two datasets considered in our analysis, DIEM-Meta [84] and LEDOV-Meta [84], are two recently proposed datasets, particularly designed to explore the performance of a dynamic saliency model under situations where understanding temporal

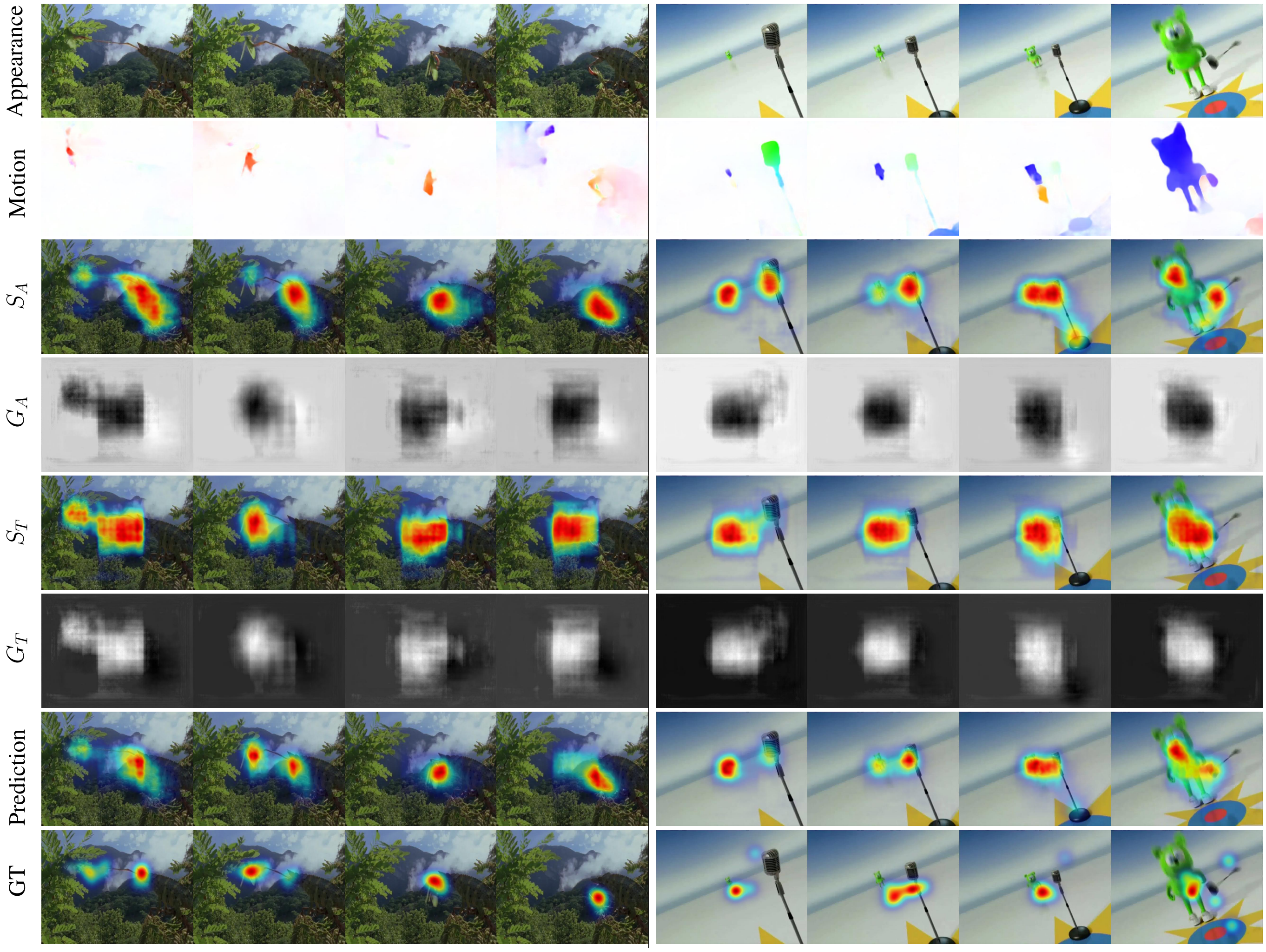


Fig. 4: Gated fusion block estimates the final saliency map by combining the appearance and the temporal maps $S_A$ and $S_T$ with the spatially varying weights $G_A$ and $G_T$.

effects is critical to give results more compatible with humans.

**UCF-Sports** dataset [81] is the smallest dataset in terms of its size, consisting of 150 videos obtained from 13 different action classes. It is originally collected for action recognition, but then enriched by [82] to include eye fixation data. The videos are annotated by 4 subjects under free-viewing condition. In the experiments, we used the same train/test splits given in [85].

**Holywood-2** dataset [82] contains 1,707 videos from Hollywood-2 action recognition dataset [86], among which 823 are used for training and the remaining 884 are left for testing. Since the videos are collected from 69 Hollywood movies with 12 action categories, its content is limited to human actions. In [82], the authors collected human fixation data for each sequence from 3 subjects under free-viewing condition. In our experiments, we use all train and test frames.

**DHF1K** [25] is the most recent and the largest video saliency dataset, which contains a total of 1000 videos with eye tracking data collected from 17 different human subjects. The authors split the dataset into 600 training, 100 validation videos and 300 test videos. The ground truth fixation data for the test split is intentionally kept hidden and the evaluation of a model on the test data is carried out by the authors themselves.

**DIEM** [83] includes 84 natural videos. Each video sequence has eye fixation data collected from approximately 50 different human subjects. Following the common experimental setup first considered in [17], we used all frames from 64 videos for training and the first 300 frames from the remaining 20 videos as test set.

**DIEM-Meta** [84] and **LEDOV-Meta** [84] are two so-called meta datasets collected from the existing video saliency datasets DIEM [83] and LEDOV [28], respectively. The main difference between these and the aforementioned datasets lies in the characteristics of the video frames they consider. They are constructed by eliminating the video frames from their original counterparts where spatial patterns are generally enough to predict where people look. To detect them, they employ a deep static saliency model that they developed. DIEM-Meta and DIEM-Meta are thus better testbeds for evaluating whether or not a dynamic saliency model learns to use the temporal domain effectively. DIEM-Meta contains only

35% of the video frames from DIEM, LEDOV-Meta includes just 20% of the original LEDOV frames.

### *B. Training Procedure*

As we mentioned previously, our network takes RGB video frames and optical flow images as inputs. We extract the frames from the videos by considering their original frame rate. We employ these RGB frames to feed our appearance stream. For the temporal stream, we generate the optical flow images between two consecutive frames by using PWC-Net [87]. We resize all the input images to $640 \times 480$ pixels and map the ground truth fixation points accordingly.

Instead of training our dynamic saliency network from scratch, we first train the subnet for the appearance stream on SALICON dataset [88]. Then, we initialize the weights of both of our subnets for spatial and temporal streams with this pre-trained static saliency model and finetune our whole two-stream network model using the dynamic saliency datasets described above. Pre-training on static data allows our dynamic saliency model to converge in fewer epochs when trained on dynamic stimuli. We use Kullback-Leibler (KL) divergence and Normalized Scanpath Saliency (NSS) loss functions (which we will explain in detail later) with Adam optimizer during the training process. We set the initial learning rate to 10e-5 and reduce it to one tenth in every 3000th iteration. The batch size is set to 8 for UCF-Sports and 16 for the other video datasets. We train our model on NVIDIA V100 GPUs (3×GPUs) and while one epoch takes approximately 2 days for the larger datasets of DHF1K, DIEM and Hollywood-2, it takes approximately 2 hours for UCF-Sports. We train our models for 2-3 epochs. Our (unoptimized) Pytorch implementation achieves a near real-time performance of 8.2 fps for frames of size $640 \times 480$ on a NVidia Tesla K40c GPU.

For our experiments on standard benchmark datasets, we consider two different training settings for dynamic stimuli. In our first setting, we use the training split of the dataset under consideration to train our proposed model. On the other hand, in our second setting, we utilize a combined training set containing training sequences from both UCF-Sports, Hollywood-2 and DHF1K datasets. The second setting further allows us to test the generalization ability of our model on DIEM, DIEM-Meta and LEDOV-Meta datasets.

**Loss functions**. In our work, we employ the combination of KL-divergence and NSS loss functions to train our proposed dynamic saliency model. As explored in previous studies, [89], [25], considering more than one loss function during training, in general, improves the model performance. Moreover, empirical experiments on the analysis of the existing automatic evaluation metrics in [90] have shown that KL-divergence and NSS are good choices for evaluating saliency models. Here, we should also note that we have one loss layer defined for the output of the merged branch. We do not define individual losses for the motion and appearance branches as we believe that they should work in harmony and complement each other in a content-dependent manner.

Let $P$ denote the predicted saliency map, $F$ represent ground truth (binary) fixation map collected from human subjects and $S$ be the ground truth (continuous) fixation density map which is generated by blurring fixation maps with a small Gaussian kernel.

KL-divergence is a widely used metric to compare two probability distributions. It has been proven to be effective for evaluating and trainig the performance of saliency models where the ground truth fixation map $S$ and the predicted saliency map $P$ are interpreted as probability distributions. Formally, KL-divergence loss function is defined as:

$$\mathcal{L}_{KL}(P,S) = \sum_i S(i) log\left(\frac{S(i)}{P(i)}\right) \ . \tag{4}$$

NSS is a location based metric which is computed as the average of the normalized predicted saliency values at fixated locations that is provided with the ground truth. By using this metric as a loss function, we force the saliency model to better detect the fixation locations and assign high likelihood scores to those pixel locations. This loss function is defined as below:

$$\mathcal{L}_{NSS}(P,F) = -\frac{1}{N}\sum_i \bar{P}(i) \times F(i) \ , \tag{5}$$

where $N$ is the total number of fixated pixels $\sum_i F(i)$ and $\bar{P}$ is the normalized saliency map $\frac{P-\mu(P)}{\sigma(P)}$.

Our final loss function is then defined as:

$$\mathcal{L}(P,F,S) = \alpha\mathcal{L}_{KL}(P,S) \ + \ \beta\mathcal{L}_{NSS}(P,F) \ , \tag{6}$$

where $\mathcal{L}_{KL}$ is the KL loss function, $\mathcal{L}_{NSS}$ is the NSS loss function, and $\alpha$ and $\beta$ are the weights for these loss functions. We first perform a set of experiments on SALICON dataset to empirically determine the optimal values of $\alpha$ and $\beta$, and then set $\alpha = 1$ and $\beta = 0.1$ for all the experiments.

### *C. Evaluation Metrics and Compared Saliency Models*

In our evaluation, we employ the following five commonly reported saliency metrics: Area Under Curve (AUC-Judd), Pearson's Correlation Coefficient (CC), Normalized Scanpath Saliency (NSS), Similarity Metric (SIM) and KL-divergence (KLDiv). For a detailed analysis of these metrics and their definitions, please refer to [90]. Each metric measures a different aspect of visual saliency and none of them is superior to the others. AUC metric considers the saliency map as classification map. A ROC curve is constituted by measuring the true and false positive rates under different binary classifier thresholds. While a score of 1 indicates a perfect match, a score close to 0.5 indicates the performance of chance. NSS is another commonly used metric, which we formally defined before while describing our loss functions. CC metric is a distribution based metric which is used to measure the linear relationship between saliency and fixation maps using the following formula:

$$CC(P,S) = \frac{\sigma(P,S)}{\sigma(P) \times \sigma(S)} \tag{7}$$

where $\sigma$ corresponds to covariance. A CC value close to +1/-1 demonstrates a perfect linear relationship. SIM is another popular metric that measures the similarity between the predicted and human saliency maps, as defined below:

$$\text{SIM}(P, S) = \sum_i \min(P_i, S_i)$$
$$\text{where } \sum_i P_i = 1 \text{ and } \sum_i S_i = 1 \tag{8}$$

KLDiv metric evaluates the dissimilatrity between two distributions. Since KLDiv represents the difference between the saliency map and the density map, a small value indicates a good result. However, we note that, according to the aforementioned study, NSS and CC seem to provide more fair results. In our experiments, we report the scores obtained with the implementations provided by MIT benchmark website[2].

We compare our method with ten different models: SalGAN [91], PQFT [10], [46], AWS-D [23], [27], OM-CNN [28], ACLNet [25], SalEMA [64], STRA-Net [66], and TASED-Net [65]. Among these, SalGAN [91] is the only static saliency model that gives the state-of-the-art results in the image datasets. We evaluate this method on video datasets considering each frame as a static image. PQFT [10], [46], and AWS-D [23] are non-deep learning models whereas all the other models employs deep learning techniques to predict where people look in videos. We note that in [27], the authors tested different fusion strategies with static weighting schemes and here we only report the results obtained with convolutional fusion strategy, which was shown to perform better than the others.

In our experiments, we use the implementations and the trained models provided by the authors and test our approach against them with the settings explained in Sec. IV-A for fair comparison. In particular, after a careful analysis, we notice that some methods do not report results on whole test set of Hollywood-2 and/or they mistakenly consider task-specific gaze data collected for UCF-Sports while generating the groundtruth fixation density maps. Hence, some of the results are different than those reported in the papers but they give a better picture of their performances. Moreover, in our experiments, we also provide the results of single-stream versions of our model that respectively consider either spatial or temporal information.

### D. Qualitative and Quantitative Results

TABLE I: Performance comparison on UCF-Sports dataset. The best and the second best performing models are shown in bold typeface and underlined, respectively.

| Method | Metric | AUC-J↑ | CC↑ | NSS↑ | SIM↑ | KLDiv↓ |
|---|---|---|---|---|---|---|
| Static | SalGAN | 0.869 | 0.389 | 2.074 | 0.258 | 2.169 |
| Dynamic | PQFT* | 0.776 | 0.211 | 1.189 | 0.157 | 2.458 |
| | Fang et al.* | 0.879 | 0.387 | 2.319 | 0.247 | 2.012 |
| | AWS-D* | 0.845 | 0.313 | 1.870 | 0.195 | 2.202 |
| | Bak et al. | 0.864 | 0.387 | 2.231 | 0.130 | 2.575 |
| | OM-CNN | 0.880 | 0.398 | 2.443 | 0.294 | 1.902 |
| | ACLNet | 0.876 | 0.367 | 2.045 | 0.292 | 2.135 |
| | SalEMA | 0.895 | 0.470 | 2.979 | **0.384** | 1.728 |
| | STRA-Net | 0.902 | 0.479 | 2.916 | **0.384** | 2.483 |
| | TASED-Net | 0.887 | 0.453 | 2.680 | 0.369 | 1.876 |
| Ours (Single) | Spatial | 0.870 | 0.461 | 3.029 | 0.377 | 2.504 |
| | Temporal | 0.851 | 0.418 | 2.535 | 0.345 | 2.721 |
| Ours (Gated) | Setting 1 | **0.914** | **0.526** | **3.333** | 0.382 | **1.516** |
| | Setting 2 | 0.911 | 0.499 | 2.980 | 0.353 | 1.568 |

* Non-deep learning model

TABLE II: Performance comparison on Hollywood-2 dataset.

| Method | Metric | AUC-J↑ | CC↑ | NSS↑ | SIM↑ | KLDiv↓ |
|---|---|---|---|---|---|---|
| Static | SalGAN | 0.892 | 0.428 | 2.383 | 0.298 | 1.760 |
| Dynamic | PQFT* | 0.689 | 0.150 | 0.610 | 0.139 | 2.387 |
| | Fang et al.* | 0.862 | 0.312 | 1.614 | 0.221 | 1.781 |
| | AWS-D* | 0.747 | 0.227 | 0.994 | 0.193 | 2.256 |
| | Bak et al. | 0.840 | 0.310 | 1.439 | 0.158 | 2.339 |
| | OM-CNN | 0.893 | 0.430 | 2.625 | 0.330 | 1.896 |
| | ACLNet | 0.899 | 0.459 | 2.463 | 0.342 | 1.701 |
| | SalEMA | 0.873 | 0.383 | 2.226 | 0.330 | 3.157 |
| | STRA-Net | 0.913 | 0.558 | 3.226 | 0.459 | 2.251 |
| | TASED-Net | 0.916 | **0.570** | **3.324** | **0.471** | 2.740 |
| Ours (Single) | Spatial | 0.904 | 0.501 | 3.051 | 0.378 | 1.473 |
| | Temporal | 0.898 | 0.489 | 2.581 | 0.362 | 1.468 |
| Ours (Gated) | Setting 1 | 0.914 | 0.549 | 3.114 | 0.413 | 1.277 |
| | Setting 2 | **0.919** | 0.563 | 3.201 | 0.424 | **1.242** |

* Non-deep learning model

**Performance on UCF-Sports.** Table I reports the comparative results on UCF-Sports test set, which contains 43 sequences. As can be seen, the single-stream versions of our proposed model gives worse scores than our full model. Moreover, spatial stream generally predicts saliency much better than the temporal stream, which is a trend that we observe on the other standard benchmark datasets too. Our model trained only on UCF-Sports outperforms all the competing models in most of the metrics. It results in a performance very close to those of SalEMA and STRA-Net in terms of SIM. We believe that weighting the predictions by the spatial and temporal streams using a gating mechanism allows the model to better handle the variations throughout video sequence, thus resulting in more accurate saliency maps on this action-specific relatively small dataset.

**Performance on Hollywood-2.** In our experiments on Hollywood-2 dataset, we use all the frames from the test set that contains 884 video sequences. In that regard, it is the largest test set that we considered in our experimental evaluation. In Table II, we provide comparison against the competing saliency models. Our results show that our model gives better saliency predictions than all the other methods in terms of the AUC-J and KLDiv metrics. The performance of the model trained considering our second training setting that includes a larger and more diverse training set provides much better results than the one trained with the first setting. In terms of the remaining evaluation metrics, our results are highly competitive as compared to the recent state-of-the-art models, namely STRA-Net and TASED-Net, as well.

**Performance on DHF1K.** We test the performance of our model on the recently proposed DHF1K video saliency dataset, which includes 300 test videos. As mentioned before, the annotations for the test split are not publicly available and all the evaluations are carried out externally by the authors of the dataset. As Table III shows, our proposed model achieves performance on par with the state-of-the-art models. In terms

[2] https://github.com/cvzoya/saliency/tree/master/code_forMetrics

TABLE III: Performance comparison on DHF1K dataset.

| Method | Metric | AUC-J↑ | CC↑ | NSS↑ | SIM↑ |
|---|---|---|---|---|---|
| Static | SalGAN | 0.866 | 0.370 | 2.043 | 0.262 |
| Dynamic | PQFT* | 0.699 | 0.137 | 0.749 | 0.139 |
| | Fang et al.* | 0.819 | 0.273 | 1.539 | 0.198 |
| | AWS-D* | 0.703 | 0.174 | 0.940 | 0.157 |
| | Bak et al. | 0.834 | 0.325 | 1.632 | 0.197 |
| | OM-CNN | 0.856 | 0.344 | 1.911 | 0.256 |
| | ACLNet | 0.890 | 0.434 | 2.354 | 0.315 |
| | SalEMA | 0.890 | 0.449 | 2.574 | **0.466** |
| | STRA-Net | **0.895** | 0.458 | 2.558 | 0.355 |
| | TASED-Net | **0.895** | **0.470** | **2.667** | 0.361 |
| Ours (Gated) | Setting 1 | 0.891 | 0.448 | 2.505 | 0.326 |
| | Setting 2 | **0.895** | 0.457 | 2.528 | 0.321 |

* Non-deep learning model

TABLE IV: Performance comparison on DIEM dataset.

| Method | Metric | AUC-J↑ | CC↑ | NSS↑ | SIM↑ | KLDiv↓ |
|---|---|---|---|---|---|---|
| Static | SalGAN | 0.860 | 0.492 | 2.068 | 0.392 | 1.431 |
| Dynamic | PQFT* | 0.680 | 0.190 | 0.656 | 0.220 | 2.140 |
| | Fang et al.* | 0.825 | 0.360 | 1.407 | 0.313 | 1.688 |
| | AWS-D* | 0.768 | 0.313 | 1.228 | 0.272 | 1.825 |
| | Bak et al. | 0.810 | 0.313 | 1.212 | 0.206 | 2.050 |
| | OM-CNN | 0.847 | 0.464 | 2.037 | 0.381 | 1.599 |
| | ACLNet | **0.878** | **0.554** | 2.283 | 0.444 | 1.331 |
| | SalEMA | 0.863 | 0.513 | 2.249 | 0.452 | 2.393 |
| | STRA-Net | 0.864 | 0.527 | 2.277 | 0.456 | 2.461 |
| | TASED-Net | 0.872 | 0.535 | 2.259 | **0.470** | 2.635 |
| Ours (Single) | Spatial | 0.868 | 0.512 | 2.202 | 0.439 | 1.387 |
| | Temporal | 0.846 | 0.446 | 1.785 | 0.391 | 1.513 |
| Ours (Gated) | Setting 1 | 0.870 | 0.543 | **2.313** | 0.454 | 1.401 |
| | Setting 2 | 0.874 | 0.525 | 2.228 | 0.421 | **1.176** |

* Non-deep learning model

TABLE V: Performance comparison on DIEM-Meta dataset.

| Method \ Metric | AUC-J↑ | CC↑ | NSS↑ | SIM↑ | KLDiv↓ |
|---|---|---|---|---|---|
| ACLNet | 0.845 | 0.437 | 1.627 | 0.391 | 1.473 |
| SalEMA | 0.832 | 0.392 | 1.576 | 0.374 | 1.664 |
| STRA-Net | 0.840 | 0.419 | 1.637 | 0.385 | 1.634 |
| TASED-Net | 0.857 | 0.455 | 1.810 | **0.416** | 1.479 |
| Ours | **0.857** | **0.460** | **1.814** | 0.395 | **1.305** |

TABLE VI: Performance comparison on LEDOV-Meta dataset.

| Method \ Metric | AUC-J↑ | CC↑ | NSS↑ | SIM↑ | KLDiv↓ |
|---|---|---|---|---|---|
| ACLNet | 0.879 | 0.384 | 1.750 | 0.342 | 1.837 |
| SalEMA | 0.863 | 0.380 | 1.815 | 0.353 | 1.850 |
| STRA-Net | **0.893** | 0.423 | 2.041 | 0.370 | 2.304 |
| TASED-Net | 0.882 | **0.489** | **2.450** | **0.403** | 1.697 |
| Ours | 0.892 | 0.457 | 2.190 | 0.370 | **1.485** |

of AUC-J, along with the recent STRA-Net and TASED-Net models, it outperforms all the other saliency models. In terms of CC, our model gives roughly the second best result.

**Performance on DIEM.** We also evaluate our model on DIEM test set consisting of 20 videos. Table IV summarizes these quantitative results. As can be seen, our model achieves the highest scores in NSS and KLDiv metrics and very competitive in others. The second setting demonstrates the generalization capability of our proposed approach as compared to the recent models like SalEMA, STRA-Net and TASED-Net.

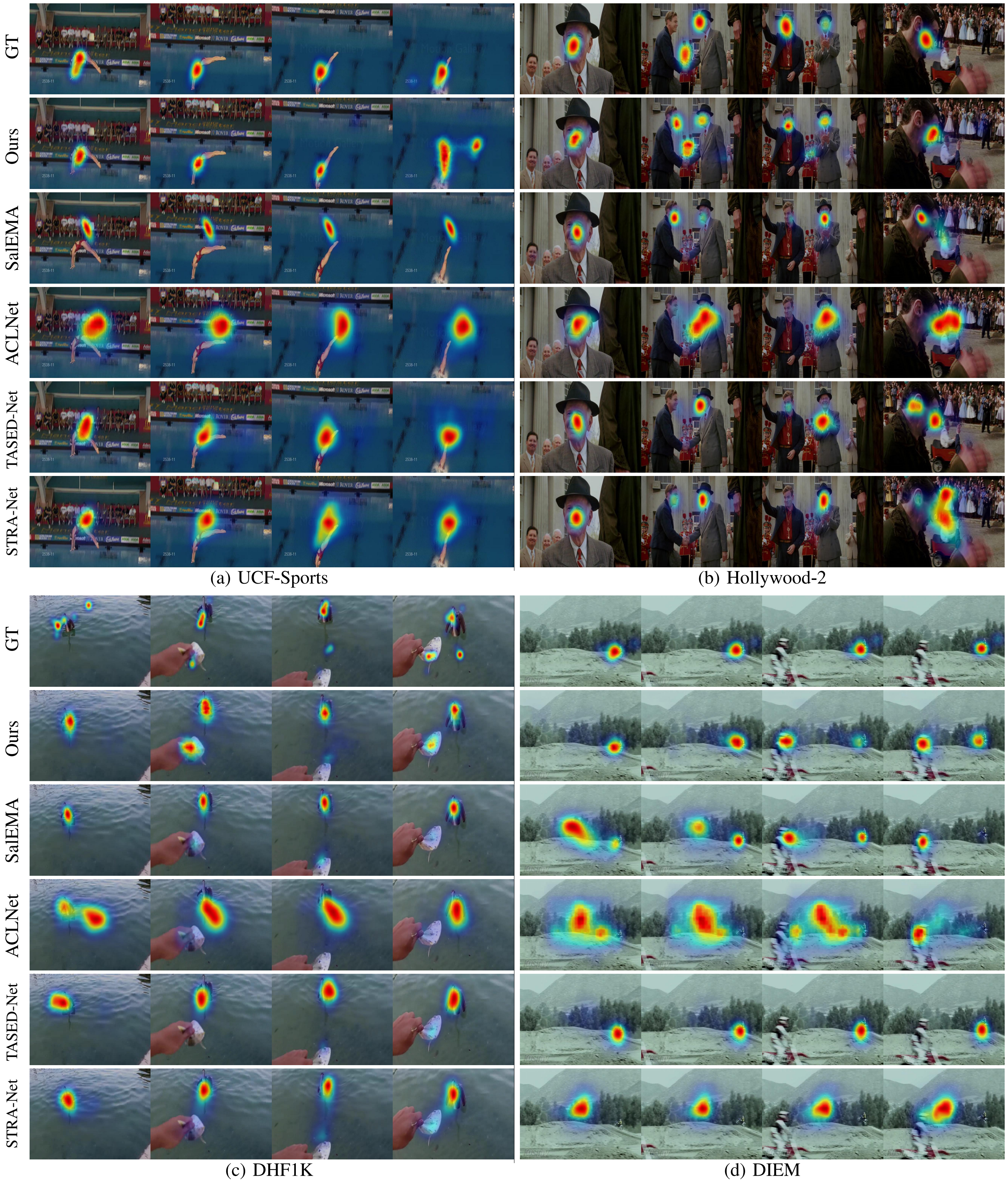
In Fig. 5, we show some sample saliency maps predicted by our proposed model and three other deep saliency networks: ACLNet, SalEMA, STRA-Net, and TASED-Net models. As one can observe, our model makes generally better predictions than the competing approaches. For instance, for the sequence from UCF-Sports (Fig. 5a) most the models fail to identify the salient region on the swimmer, or for the sequence from the Hollywood-2 dataset (Fig. 5b) our model is the only model that correctly predicts the soldier at the center of the background as salient. Similar kind of observations are also valid for the sample sequences from DHF1K (Fig. 5c) and DIEM (Fig. 5d) datasets.

**Performance on DIEM-Meta and LEDOV-Meta.** As mentioned before, [84] have recently showed that most of the current benchmarks for video saliency include many sequences in which spatial attention is more dominant than temporal effects in describing saliency. DIEM-Meta and LEDOV-Meta datasets are curated in a special way to contain video frames in which temporal signals are found to be more influential than appearance cues. Hence, they both offer a better way to test how well a dynamic saliency model utilizes temporal information. In our experimental evaluation, we compare our proposed model with the state-of-the-art deep saliency models, which are all trained on the combined training set that includes frames from DIEM or LEDOV datasets. As can be seen from Table V and Table VI, our model outperforms all the other models in DIEM-Meta, and is the second best model in LEDOV-Meta, achieving highly competitive performances. These results demonstrate the effectiveness of the proposed gated mechanism and its ability to use temporal information to the full extent, as compared to the state-of-the-art approaches.

Overall, the results reported on all the six datasets used in our experimental analysis suggest that our model has better capacity to mimic human attention mechanism by combining the temporal and static clues in an effective way. It has a better generalization ability that it can predict where people look at the videos from unseen domains much better. Moreover, it utilizes the temporal information more successfully with its gated fusion mechanism, which adaptively integrates spatial and temporal cues depending on video content.

### E. Ablation study.

In this section, we aim to analyze the influence of each component of our proposed deep dynamic saliency model. We perform the ablation study on UCF-Sports, DIEM-Meta, LEDOV-Meta datasets by disabling or removing some blocks of our model and by examining how these changes affect the model performance. As done in training our proposed model, for each version of our model under evaluation, we first train a single stream model on SALICON dataset and then use it to finetune the actual two-stream version on UCF-Sports dataset. Table VII shows the contributions of different components of our saliency model on UCF-Sports dataset. Moreover, to demonstrate the generalization capabilities of each version of our model, in Table VIII and Table IX, we evaluate their performance on LEDOV-Meta and DIEM-Meta datasets, respectively. In the following, we summarize our observations.



(a) UCF-Sports

(b) Hollywood-2

(c) DHF1K

(d) DIEM

Fig. 5: Qualitative results of our proposed framework and the deep learning based SalEMA, ACLNet and SalGAN models. Our approach, in general, produces more accurate saliency predictions than these state-of-the-art models.

**Effect of gated fusion.** As we emphasized before, the role of gated fusion block is to adaptively integrate spatial and temporal streams is a key component of our model. In our analysis, we replace the gated fusion block with a standard $1\times1$ convolution layer (that version of our model is referred to

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2021.3094974, IEEE Transactions on Cognitive and Developmental Systems

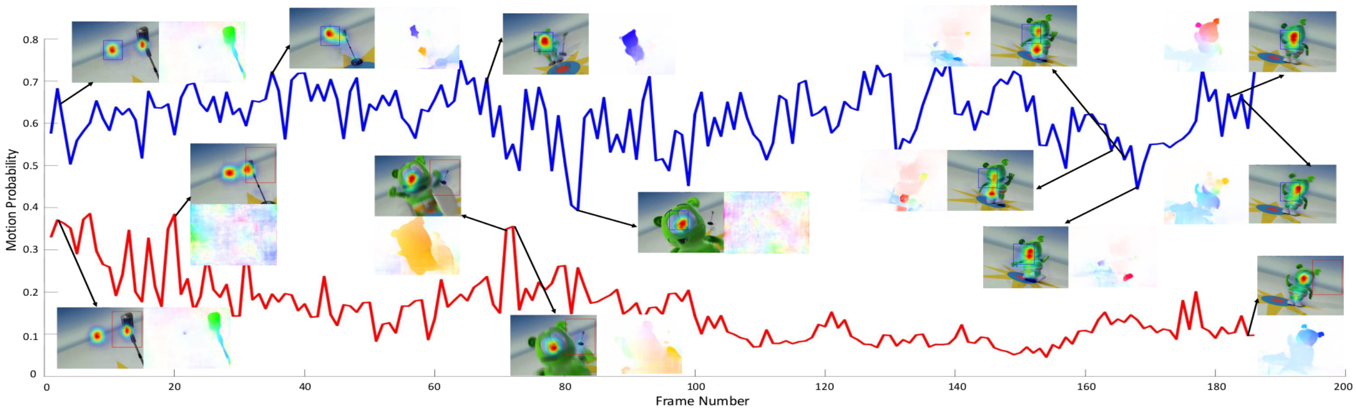


Fig. 6: Our model dynamically decides the contribution of motion and appearance streams via gated fusion. Here, we plot the average motion probabilities (the contribution of motion stream) for two regions having different characteristic, one containing a moving object (the gummy bear) and the other with relatively no motion, shown with red and blue, respectively. As can be seen, our model assigns higher weights to the motion stream when motion becomes the dominant visual cue, and the weights adaptively change throughout the sequence.

as "w/o gated fusion")[3]. As can be seen from Table VII-IX, the performance of the model decreases considerably without the gated fusion mechanism. That is, using a dynamic weighting strategy, instead of a fixed weighting scheme (learned via $1 \times 1$ convolution), generates much better predictions. Fig. 6 shows a visualization of how our proposed gated fusion operates in an adaptive manner, demonstrating the behavior of the weighting scheme for both static and dynamic parts of a given video. In particular, we plot the motion probabilities averaged within the corresponding image regions over time, which clearly shows that the motion probability (the contribution of motion stream) for the region that contains a moving object is, in general, much higher than that of the static region. Moreover, depending on the characteristics of the regions, it shows the changes in the motion probabilities throughout the whole sequence. For example, when no motion is taking place in the region initially containing the moving object, the weight of the temporal stream starts to fall. These results supports our main claim that the proposed gated fusion mechanism successfully adapts itself according to the content of the video, as opposed to having a fixed fusion strategy as in the competing approaches.

**Effect of multi-level information.** Previous studies demonstrate that low and high-level cues are equally important for saliency prediction [8], [9]. Motivated with these, we included a multi-level information block to fuse features extracted from different levels of our deep model. For this analysis, we disable this multi-level information block and train a single-scale model instead. Compared to our full model, disabling this block reduces the performance as can be seen in Table VII-IX. Employing a representation that contains information from low and high levels helps to improve the performance of our model. We speculate that our multi-level information block allows the network to better identify the regions semantically important for saliency.

TABLE VII: Ablation study on UCF-Sports dataset.

| Method \ Metric | AUC-J↑ | CC↑ | NSS↑ | SIM↑ | KLDiv↓ |
|---|---|---|---|---|---|
| w/o spatial attention | 0.872 | 0.474 | 2.884 | 0.374 | 2.223 |
| w/o channel-wise attention | 0.892 | 0.489 | 2.923 | 0.319 | 1.707 |
| w/o spatial & ch.-wise attention | 0.875 | 0.447 | 2.885 | 0.364 | 2.646 |
| w/o multi-level information | 0.890 | 0.484 | 2.755 | 0.303 | 1.711 |
| w/o gated fusion | 0.900 | 0.480 | 2.913 | 0.353 | 1.676 |
| full model | **0.914** | **0.526** | **3.333** | **0.382** | **1.516** |

TABLE VIII: Ablation study on LEDOV-Meta dataset.

| Method \ Metric | AUC-J↑ | CC↑ | NSS↑ | SIM↑ | KLDiv↓ |
|---|---|---|---|---|---|
| w/o spatial attention | 0.859 | 0.380 | 1.861 | 0.339 | 2.091 |
| w/o channel-wise attention | 0.884 | 0.420 | 1.997 | 0.318 | 1.589 |
| w/o spatial & ch.-wise attention | 0.820 | 0.310 | 1.487 | 0.297 | 2.906 |
| w/o multi-level information | **0.895** | **0.458** | 2.074 | 0.329 | 1.517 |
| w/o gated fusion | 0.852 | 0.381 | 1.743 | 0.280 | 1.765 |
| full model | 0.893 | 0.441 | **2.123** | **0.356** | **1.483** |

TABLE IX: Ablation study on DIEM-Meta dataset.

| Method \ Metric | AUC-J↑ | CC↑ | NSS↑ | SIM↑ | KLDiv↓ |
|---|---|---|---|---|---|
| w/o spatial attention | 0.806 | 0.338 | 1.372 | 0.334 | 2.155 |
| w/o channel-wise attention | 0.823 | **0.387** | 1.527 | 0.330 | **1.489** |
| w/o spatial & ch.-wise attention | 0.758 | 0.251 | 1.008 | 0.268 | 3.592 |
| w/o multi-level information | 0.809 | 0.370 | 1.428 | 0.314 | 1.567 |
| w/o gated fusion | 0.800 | 0.359 | 1.373 | 0.304 | 1.620 |
| full model | **0.827** | 0.380 | **1.531** | **0.345** | 1.511 |

**Effect of attention blocks.** As discussed before, the reasons we introduce the attention blocks are to eliminate the irrelevant features via the spatial attention and to choose the most informative feature channels via the channel-wise attention when processing a video frame. In this experiment, we remove the spatial and the channel-wise attention blocks from our full model and train two different models, respectively. The results given in Table VII support our assertion that both of these attention blocks improve the model performance. Disabling them results in a much lower performance as compared to that of the full model.

[3] Other fusion strategies such as average and max fusion were investigated in [27] and shown to be less effective than convolution fusion. Hence, we did not consider them in our ablation study.

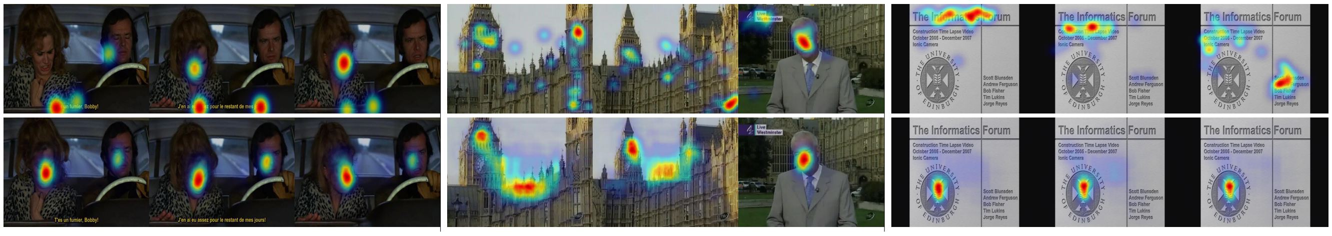


Fig. 7: Sample failure cases. Our model performs poorly on videos that contain readable text or large objects with fine details. The first shortcoming is inevitable since the data seen during training lack enough number of samples to learn to mimic eye gaze movement during reading effectively. The second drawback, on the other hand, can be attributed to the underlying convolutional neural architecture that our model depends on.

## V. Summary and Conclusion

In this study, we proposed a new spatio-temporal saliency network for video saliency. It follows a two-stream network architecture that processes spatial and temporal information in separate streams, but it extends the standard structure in many ways. First, it includes a gated fusion block that performs integration of spatial and temporal streams in a more dynamic manner by deciding the contribution of each channel one frame at a time. Second, it utilizes a multi-level information block that allows for performing multi-scale processing of appearance and motion features. Finally, it employs spatial and channel-wise attention blocks to further increase the selectivity. Our extensive set of experiments on six different benchmark datasets shows the effectiveness of the proposed model in extracting the most salient parts of the video frames both qualitatively and quantitatively. Moreover, our ablation study demonstrates the gains achieved by each component of our model. Our analysis reveals that the proposed model deals with the videos from unseen domains much better that the existing dynamic saliency models. Additionally, it uses temporal cues more effectively via the proposed gated fusion mechanism which allows for adaptive integration of spatial and temporal streams.

As can be seen in Fig. 7 our model performs poorly especially for the videos containing readable text and repetitive patterns that cover most of the frames. Since our model is not able to explicitly interpret text from semantically, it can not mimic the reading behaviour of the human. Moreover, exploring the details in the objects that have repetitive patterns is particularly challenging for the models that are based on convolutional neural networks due to the effective receptive fields of the learned filters.

We believe that our work highlights several important directions to pursue for better modeling of saliency in videos. As future work, we plan to explore more efficient ways to include the temporal information. For instance, instead of using optical flow images, one can use features extracted from early and mid layers of an optical flow network model to encode motion information. This can reduce the memory footprint of the model and decreases the running times. Another interesting research direction is to adapt the proposed gating mechanism for an architecture that alternatively utilizes 3D convolutions instead of a two-stream framework.

## Acknowledgments

This work was supported in part by GEBIP 2018 Award of the Turkish Academy of Sciences to E. Erdem, BAGEP 2021 Award of the Science Academy to A. Erdem.

## References

[1] L. Itti, C. Koch, and E. Niebur, “A model of saliency-based visual attention for rapid scene analysis,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[2] A. Borji and L. Itti, “State-of-the-art in visual attention modeling,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 185–207, 2013.

[3] H. Kim and S. Lee, “Transition of visual attention assessment in stereoscopic images with evaluation of subjective visual quality and discomfort,” *IEEE Transactions on Multimedia*, vol. 17, no. 12, pp. 2198–2209, 2015.

[4] K. Gu, S. Wang, H. Yang, W. Lin, G. Zhai, X. Yang, and W. Zhang, “Saliency-guided quality assessment of screen content images,” *IEEE Transactions on Multimedia*, vol. 18, no. 6, pp. 1098–1110, 2016.

[5] Y. Fang, Z. Chen, W. Lin, and C. Lin, “Saliency detection in the compressed domain for adaptive image retargeting,” *IEEE Transactions on Image Processing*, vol. 21, no. 9, pp. 3888–3901, 2012.

[6] D. Chen and Y. Luo, “Preserving motion-tolerant contextual visual saliency for video resizing,” *IEEE Transactions on Multimedia*, vol. 15, no. 7, pp. 1616–1627, 2013.

[7] G. Evangelopoulos, A. Zlatintsi, A. Potamianos, P. Maragos, K. Rapantzikos, G. Skoumas, and Y. Avrithis, “Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention,” *IEEE Transactions on Multimedia*, vol. 15, no. 7, pp. 1553–1568, 2013.

[8] N. D. B. Bruce, C. Catton, and S. Janjic, “A deeper look at saliency: Feature contrast, semantics, and beyond,” in *CVPR*, 2016, pp. 516–524.

[9] Z. Bylinskii, A. Recasens, A. Borji, A. Oliva, A. Torralba, and F. Durand, “Where should saliency models look next?” in *Proc. ECCV*, 2016, pp. 809–824.

[10] C. Guo, Q. Ma, and L. Zhang, “Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform,” in *Proc. CVPR*, 2008, pp. 1–8.

[11] X. Cui, Q. Liu, and D. Metaxas, “Temporal spectral residual: fast motion saliency detection,” in *ACM MM*, 2009, pp. 617–620.

[12] H. J. Seo and P. Milanfar, “Static and space-time visual saliency detection by self-resemblance,” *Journal of Vision*, vol. 9, no. 12, pp. 15–15, 2009.

[13] W. Sultani and I. Saleemi, “Human action recognition across datasets by foreground-weighted histogram decomposition,” in *Proc. CVPR*, 2014, pp. 764–771.

[14] T. Mauthner, H. Possegger, G. Waltner, and H. Bischof, “Encoding based saliency detection for videos and images,” in *CVPR*, 2015, pp. 2494–2502.

[15] Y. Luo and Q. Tian, “Spatio-temporal enhanced sparse feature selection for video saliency estimation,” in *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, June 2012, pp. 33–38.

[16] S. Mathe and C. Sminchisescu, “Dynamic eye movement datasets and learnt saliency models for visual action recognition,” in *Proc. ECCV*, 2012, pp. 842–856.

[17] D. Rudoy, D. B. Goldman, E. Shechtman, and L. Zelnik-Manor, "Learning video saliency from human gaze using candidate selection," in *Proc. CVPR*, 2013, pp. 1147–1154.

[18] S. Zhong, Y. Liu, F. Ren, J. Zhang, and T. Ren, "Video saliency detection via dynamic consistent spatio-temporal attention modelling," in *Proc., AAAI*, 2013.

[19] Z. Liu, X. Zhang, S. Luo, and O. Le Meur, "Superpixel-based spatiotemporal saliency detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 9, pp. 1522–1540, 2014.

[20] F. Zhou, S. B. Kang, and M. F. Cohen, "Time-mapping using space-time saliency," in *Proc. CVPR*, 2014, pp. 3358–3365.

[21] J. Zhao, C. Siagian, and L. Itti, "Fixation bank: Learning to reweight fixation candidates," in *CVPR*, 2015, pp. 3174–3182.

[22] S. H. Khatoonabadi, N. Vasconcelos, I. V. Bajić, and Yufeng Shan, "How many bits does it take for a stimulus to be salient?" in *Proc. CVPR*, 2015, pp. 5501–5510.

[23] V. Leborán, A. García-Díaz, X. R. Fdez-Vidal, and X. M. Pardo, "Dynamic whitening saliency," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 5, pp. 893–907, 2017.

[24] L. Bazzani, H. Larochelle, and L. Torresani, "Recurrent mixture density network for spatiotemporal visual attention," in *Proc. ICLR*, 2017.

[25] W. Wang, J. Shen, J. Xie, M. Cheng, H. Ling, and A. Borji, "Revisiting video saliency prediction in the deep learning era," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[26] A. Palazzi, D. Abati, S. Calderara, F. Solera, and R. Cucchiara, "Predicting the driver's focus of attention: the dr(eye)ve project," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[27] 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, 2018.

[28] L. Jiang, M. Xu, T. Liu, M. Qiao, and Z. Wang, "Deepvs: A deep learning based video saliency prediction approach," in *Proc. ECCV*, 2018, pp. 625–642.

[29] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *CVPR*, 2014, pp. 1725–1732.

[30] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS)*, 2014, p. 568–576.

[31] J. Wolfe, M. Võ, K. K. Evans, and M. R. Greene, "Visual search in scenes involves selective and nonselective pathways," *Trends in Cognitive Sciences*, vol. 15, pp. 77–84, 2011.

[32] A. C. R. Farivar, O. Blanke, "Dorsal-ventral integration in the recognition of motion-defined unfamiliar faces," *J Neurosci*, vol. 29, 2009.

[33] M. A. Goodale and A. D. Milner, "Separate visual pathways for perception and action," *Trends in Neurosciences*, vol. 15, no. 1, pp. 20–?25, 1992.

[34] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive Psychology*, vol. 12, no. 1, pp. 97–136, 1980.

[35] C. Koch and S. Ullman, "Shifts in selective visual attention: Towards the underlying neural circuitry," *Human neurobiology*, vol. 4, pp. 219–27, 1985.

[36] X. Huang, C. Shen, X. Boix, and Q. Zhao, "SALICON: Reducing the semantic gap in saliency prediction by adapting deep neural networks," in *Proc. ICCV*, 2015, pp. 262–270.

[37] S. Jetley, N. Murray, and E. Vig, "End-to-end saliency mapping via probability distribution prediction," in *Proc. CVPR*, 2016, pp. 5753–5761.

[38] S. S. S. Kruthiventi, K. Ayush, and R. V. Babu, "Deepfix: A fully convolutional neural network for predicting human eye fixations," *IEEE Transactions on Image Processing*, vol. 26, no. 9, pp. 4446–4456, 2017.

[39] N. Liu, J. Han, T. Liu, and X. Li, "Learning to predict eye fixations via multiresolution convolutional neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 2, pp. 392–404, 2018.

[40] J. Pan, E. Sayrol, X. Giró-i Nieto, K. McGuinness, and N. E. OConnor, "Shallow and deep convolutional networks for saliency prediction," in *Proc. CVPR*, 2016, pp. 598–606.

[41] W. Wang and J. Shen, "Deep visual attention prediction," *IEEE Transactions on Image Processing*, vol. 27, no. 5, pp. 2368–2378, 2018.

[42] E. Vig, M. Dorr, and D. Cox, "Large-scale optimization of hierarchical features for saliency prediction in natural images," in *Proc. CVPR*, 2014, pp. 2798–2805.

[43] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, "Predicting human eye fixations via an lstm-based saliency attentive model," *IEEE Transactions on Image Processing*, vol. 27, no. 10, pp. 5142–5154, 2018.

[44] Z. Wang, Z. Liu, W. Wei, and H. Duan, "Saled: Saliency prediction with a pithy encoder-decoder architecture sensing local and global information," *Image and Vision Computing*, vol. 109, p. 104149, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0262885621000548

[45] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Proceedings of the 19th International Conference on Neural Information Processing Systems (NIPS)*, 2006, pp. 545–552.

[46] Y. Fang, Z. Wang, W. Lin, and Z. Fang, "Video saliency incorporating spatiotemporal cues and uncertainty weighting," *IEEE Transactions on Image Processing*, vol. 23, no. 9, pp. 3910–3921, 2014.

[47] E. Rahtu, J. Kannala, M. Salo, and J. Heikkilä, "Segmenting salient objects from images and videos," in *Proc. ECCV*, 2010, pp. 366–379.

[48] Z. Liu, J. Li, L. Ye, G. Sun, and L. Shen, "Saliency detection for unconstrained videos using superpixel-level graph and spatiotemporal propagation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 12, pp. 2527–2542, 2017.

[49] W. Wang, J. Shen, and L. Shao, "Consistent video saliency using local gradient flow optimization and global refinement," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 4185–4196, 2015.

[50] H. Kim, Y. Kim, J. Sim, and C. Kim, "Spatiotemporal saliency detection for video sequences based on random walk with restart," *IEEE Transactions on Image Processing*, vol. 24, no. 8, pp. 2552–2564, 2015.

[51] J. Li, Z. Liu, X. Zhang, O. Le Meur, and L. Shen, "Spatiotemporal Saliency Detection Based on Superpixel-level Trajectory," *Signal Processing: Image Communication*, vol. 38, pp. 100–114, 2015.

[52] W. Wang, J. Shen, and F. Porikli, "Saliency-aware geodesic video object segmentation," in *Proc. CVPR*, 2015, pp. 3395–3402.

[53] X. Zhou, Z. Liu, C. Gong, and W. Liu, "Improving video saliency detection via localized estimation and spatiotemporal refinement," *IEEE Transactions on Multimedia*, vol. 20, no. 11, pp. 2993–3007, 2018.

[54] G. Li and Y. Yu, "Visual saliency based on multiscale deep features," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 5455–5463.

[55] Z. Luo, A. Mishra, A. Achkar, J. Eichel, S. Li, and P.-M. Jodoin, "Non-local deep features for salient object detection," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6593–6601.

[56] L. Wang, H. Lu, X. Ruan, and M.-H. Yang, "Deep networks for saliency detection via local estimation and global search," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3183–3192.

[57] P. Zhang, D. Wang, H. Lu, H. Wang, and X. Ruan, "Amulet: Aggregating multi-level convolutional features for salient object detection," 2017.

[58] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. H. S. Torr, "Deeply supervised salient object detection with short connections," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 4, p. 815–828, Apr 2019.

[59] H. Song, W. Wang, S. Zhao, J. Shen, and K.-M. Lam, "Pyramid dilated deeper convlstm for video salient object detection," in *CECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham: Springer International Publishing, 2018, pp. 744–760.

[60] N. Liu and J. Han, "Dhsnet: Deep hierarchical saliency network for salient object detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 678–686.

[61] T. Wang, A. Borji, L. Zhang, P. Zhang, and H. Lu, "A stagewise refinement model for detecting salient objects in images," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 4039–4048.

[62] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proc. ICCV*, 2015, pp. 4489–4497.

[63] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *Proc. ICLR*, 2016.

[64] P. Linardos, E. Mohedano, J. J. Nieto, K. McGuinness, X. Giro-i Nieto, and N. E. O'Connor, "Simple vs complex temporal recurrences for video saliency prediction," in *Proc. BMVC*, 2019.

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

[66] Q. Lai, W. Wang, H. Sun, and J. Shen, "Video saliency prediction using spatiotemporal residual attentive networks," *IEEE Trans. on Image Processing*, 2019.

[67] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016, pp. 770–778.

[68] T. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. CVPR*, 2017, pp. 936–944.

[69] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*, 2015, pp. 234–241.

[70] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. CVPR*, 2015, pp. 3431–3440.

[71] P. O. Pinheiro, T.-Y. Lin, R. Collobert, and P. Dollár, "Learning to refine object segments," in *Proc. ECCV*, 2016, pp. 75–91.

[72] T. Zhao and X. Wu, "Pyramid feature attention network for saliency detection," in *Proc. CVPR*, 2019, pp. 3080–3089.

[73] S. Dong, Z. Gao, S. Sun, X. Wang, M. M. Li, H. Zhang, G. Yang, H. Liu, and S. Li, "Holistic and deep feature pyramids for saliency detection," in *Proc. BMVC*, 2018.

[74] V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu, "Recurrent models of visual attention," in *Proc. NIPS*, 2014.

[75] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. ICML*, 2015.

[76] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang, "Multi-context attention for human pose estimation," in *Proc. CVPR*, 2017, pp. 5669–5678.

[77] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua, "SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning," in *Proc. CVPR*, 2017.

[78] W. Ren, L. Ma, J. Zhang, J. Pan, X. Cao, W. Liu, and M.-H. Yang, "Gated fusion network for single image dehazing," in *Proc. CVPR*, 2018.

[79] X. Zhang, H. Dong, Z. Hu, W.-S. Lai, F. Wang, and M.-H. Yang, "Gated fusion network for joint image deblurring and super-resolution," in *Proc. BMVC*, 2018.

[80] Y. Cheng, R. Cai, Z. Li, X. Zhao, and K. Huang, "Locality-sensitive deconvolution networks with gated fusion for RGB-D indoor semantic segmentation," in *Proc. CVPR*, 2017, pp. 1475–1483.

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

[82] S. Mathe and C. Sminchisescu, "Actions in the eye: Dynamic gaze datasets and learnt saliency models for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 7, pp. 1408–1424, 2015.

[83] 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.

[84] M. Tangemann, M. Kümmerer, T. S. Wallis, and M. Bethge, "Measuring the importance of temporal features in video saliency," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[85] T. Lan, Y. Wang, and G. Mori, "Discriminative figure-centric models for joint action localization and recognition," in *Proc. ICCV*, 2011, pp. 2003–2010.

[86] M. Marszalek, I. Laptev, and C. Schmid, "Actions in context," in *Proc. CVPR*, 2009, pp. 2929–2936.

[87] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, "PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume," in *Proc. CVPR*, 2018.

[88] M. Jiang, S. Huang, J. Duan, and Q. Zhao, "SALICON: Saliency in context," in *Proc. CVPR*, 2015, pp. 1072–1080.

[89] X. Huang, C. Shen, X. Boix, and Q. Zhao, "SALICON: Reducing the semantic gap in saliency prediction by adapting deep neural networks," in *Proc. ICCV*, 2015, pp. 262–270.

[90] 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, 2019.

[91] J. Pan, C. Canton, K. McGuinness, N. E. O'Connor, J. Torres, E. Sayrol, and X. a. Giro-i Nieto, "SalGAN: Visual saliency prediction with generative adversarial networks," in *arXiv*, 2017.

**Aysun Kocak** received her B.Sc. degree in Mathematics from Hacettepe University, Ankara, Turkey, in 2012. She is currently a Ph.D. student in the Department of Computer Engineering at Hacettepe University, Ankara, Turkey. Her research interests include machine learning, computer vision and visual saliency prediction.

**Erkut Erdem** received his Ph.D. degree from Middle East Technical University in 2008. After completing his Ph.D., he continued his post-doctoral studies with Télécom ParisTech, École Nationale Supérieure des Télécommunications, France, from 2009 to 2010. He has been an Associate Professor with the Department of Computer Engineering, Hacettepe University, Turkey, since 2014. His research interests include semantic image editing, visual saliency prediction, and integrated vision and language applications.

**Aykut Erdem** is an Associate Professor of Computer Science at Koç University. He received his Ph.D. degree from Middle East Technical University in 2008. He was a post-doctoral researcher at the Ca'Foscari University in Venice in the EU-FP7 SIMBAD project, from 2008 to 2010. Previously, he was with the Computer Engineering Department at Hacettepe University, where he was one of the directors of the Computer Vision Lab. The broad goal of his research is to explore better ways to understand, interpret and manipulate visual data. His current research focuses on investigating learning-based approaches to image editing, visual saliency estimation, and connecting vision and language.