# Interactive Video Saliency Prediction: The Stacked-convLSTM Approach

N. Wondimu[1,4][a], U. Visser[3][b] and C. Buche[1,2][c]

[1] *Lab-STICC, Brest National School of Engineering, 29280, Plouzané, France*

[2] *IRL CROSSING, CNRS, Adelaide, Australia*

[3] *University of Miami, Florida, U.S.A.*

[4] *School of Information Technology and Engineering, Addis Ababa University, Addis Ababa, Ethiopia*

Keywords: Saliency Prediction, Video Saliency, Human Attention, Gaze Prediction, ConvLSTM, Video Saliency Dataset.

Abstract: Cognitive and neuroscience of attention researches suggest the use of spatio-temporal features for an efficient video saliency prediction. This is due to the representative nature of spatio-temporal features for data collected across space and time, such as videos. Video saliency prediction aims to find visually salient regions in a stream of images. Many video saliency prediction models are proposed in the past couple of years. Due to the unique nature of videos from that of static images, the earliest efforts to employ static image saliency prediction models for video saliency prediction task yield reduced performance. Consequently, dynamic video saliency prediction models that use spatio-temporal features were introduced. These models, especially deep learning based video saliency prediction models, transformed the state-of-the-art of video saliency prediction to a better level. However, video saliency prediction still remains a considerable challenge. This has been mainly due to the complex nature of video saliency prediction and scarcity of representative saliency benchmarks. Given the importance of saliency identification for various computer vision tasks, revising and enhancing the performance of video saliency prediction models is crucial. To this end, we propose a novel interactive video saliency prediction model that employs stacked-ConvLSTM based architecture along with a novel XY-shift frame differencing custom layer. Specifically, we introduce an encoder-decoder based architecture with a prior layer undertaking XY-shift frame differencing, a residual layer fusing spatially processed (VGG-16 based) features with XY-shift frame differenced frames, and a stacked-ConvLSTM component. Extensive experimental results over the largest video saliency dataset, DHF1K, show the competitive performance of our model against the state-of-the-art models.

## 1 INTRODUCTION

It is crucial that robotic systems employ robust computational models that irreproachably mimic human's perceptive and action intelligence, in real-time. Saliency prediction is among the most significant capabilities of human visual system. The human visual system is able to quickly distinguish important scenes in its visual field. The ability to computationally model this feature of human enables efficient and realistic human-robot interaction in social standard robotic environment (Ferreira and Dias, 2014; Schillaci et al., 2013; Diaz et al., 2019). Specifically, it plays a vital role in enabling intuitive and natural human-robot interaction by letting the robot to continuously pay attention to salient regions in its visual field (Schillaci et al., 2013; Butko et al., 2008). Besides, these computational models can be used as a source of efficiency in various computer vision tasks (Zhang et al., 2018).

Saliency prediction systems have been applied to various problem domains, such as video segmentation (Fukuchi et al., 2009; Zhang et al., 2018), video captioning (Chen et al., 2018; Wang et al., 2018a), video compression (Guo and Zhang, 2009), image captioning (Cornia et al., 2018) autonomous driving (Pal et al., 2020; Lateef et al., 2021), human-robotic interaction (Schillaci et al., 2013; Schauerte and Stiefelhagen, 2014), robot navigation (Roberts et al., 2012; Chang et al., 2010), surveillance (Yubing et al., 2011;

[a] https://orcid.org/0000-0002-0726-9892

[b] https://orcid.org/0000-0002-1254-2566

[c] https://orcid.org/0000-0003-0264-2683

Wondimu, N., Visser, U. and Buche, C.
Interactive Video Saliency Prediction: The Stacked-convLSTM Approach.
DOI: 10.5220/0011664600003393
In *Proceedings of the 15th International Conference on Agents and Artificial Intelligence (ICAART 2023) - Volume 2*, pages 157-168
ISBN: 978-989-758-623-1; ISSN: 2184-433X

Shao et al., 2019), and other areas (Yun et al., 2019; Ji et al., 2022).

Visual saliency has been studied from the spatial (Shi et al., 2015a; Xie and Lu, 2011) and spatio-temporal perspectives (Marat et al., 2009) . Spatial information of individual images or frames has been used to build the earliest static image saliency prediction computational models. Several experiments also show that, computational models, especially those inspired by deep neural networks (DNN), suffice the problem of static saliency prediction (Itti et al., 1998; Harel et al., 2006; Huang et al., 2015; Wang and Shen, 2017; Pan et al., 2016). However, because of the spatio-temporal or dynamic nature of videos, almost all static image saliency prediction models show hampered performance when employed on video stimulus.

To this end, recent video saliency prediction models are considering spatio-temporal aspects of video saliency dataset. This is mainly due to the recent cognitive and neuroscience of attention research findings, asserting to the importance of spatio-temporal features for data collected across space and time (Bohic and Abraira, 2022; Amso and Scerif, 2015). Besides, advances in deep neural networks and their ability to efficiently handle spatio-temporal data contributed a lot to the growth of DNN inspired dynamic saliency prediction models.

A number of video saliency computational models have been produced in recent years. However, most models use datasets that lack generic, representative, and diverse instances in unconstrained task-independent scenarios. This has been exposing them for over-fitting (Rice et al., 2020) and incapability to work on real and diverse environment.

Very few computational models have been using diverse and representative datasets, like DHF1K (Wang et al., 2018b). The use of large and representative video saliency dataset along with advanced deep neural networks show significant performance improvement (Bak et al., 2017; Wang et al., 2018b). However,video saliency prediction problem in a complex and dynamic environment remains a challenge to this date. To this end, we propose a novel interactive stacked-ConvLSTM based video saliency model. Our architecture introduce a new XY-Shift frame differencing custom layer to boost temporal features on spatial domain. Moreover, we introduce a novel way of fusing temporally magnified spatio temporal features with features engineered with spatial feature extractors like VGG-16 (Simonyan and Zisserman, 2014). We use stacked-ConvLSTM component (Shi et al., 2015b) for sequential fixation prediction over successive frames. A successive experiments we conducted on the largest video saliency dataset,DHF1K (Wang et al., 2018b), show that our model achieve a competitive result against the state-of-the-art methods.

The rest of the paper is organized as follows. The second part briefly introduces related research works, the third part introduce the saliency prediction model proposed in this paper in detail, the fourth part shows experimental details of this paper, and finally, a summary of this paper is presented.

## 2 RELATED WORKS

Recent researches on visual saliency have been consecutively redefining the state-of-the-art in the area. Most of the earliest saliency models are constructed from still images. These computational models assume that conspicuous visual features “pop-out” and involuntarily capture attention (Borji and Itti, 2012). However, the performance of these models is significantly hampered as it belittles the impact of temporal features. To this end, recent advances on visual saliency prediction consider dynamic features for visual saliency prediction. The growth in this field of saliency is due to the growth in the area of deep learning and the availability of larger video saliency datasets. In this section, existing visual saliency prediction models that define the state-of-the-art in the area are briefly reviewed.

### 2.1 Saliency Models

Researches on human gaze fixation prediction or video saliency prediction is dating back to (Itti and Koch, 2001; Itti et al., 1998). The earliest saliency prediction methods are based on various low-level manual features of still image, such as color contrast, edge, center prior and orientation to produce a “saliency map” (Harel et al., 2006; Le Meur et al., 2006; Bruce and Tsotsos, 2005; Judd et al., 2009; Wang et al., 2016; Yang et al., 2013; Jiang et al., 2013). A saliency map is an image that highlights the region on which human gaze could focus on a various probabilistic level.

Low-level feature based saliency models can work robustly on the simplest detection tasks. However, these models fail to perform well on a more complex image structures. To this end, various deep learning based static saliency researches are published Hou et al. (Hou et al., 2017), Lee et al. (Itti and Koch, 2001) and Li and Yu (Itti et al., 1998) Wang et al. (Wang et al., 2017a) and Zhang et al. (Zhang et al., 2017) (Vig et al., 2014; Kruthiventi et al., 2017; Huang et al., 2015; Liu et al., 2016; Pan et al., 2016; Wang and Shen, 2017). These models have achieved a remarkable result using the powerful learning ability of neural networks and growth in the size and quality of visual saliency datasets (Huang et al., 2015).

Static image saliency research is almost mature. However, subsequent trials to employ these models on video show a reduced performance (Mahadevan and Vasconcelos, 2009). These is mainly due to the frequent change in salient-goal over time in a sequence of frames. Furthermore, convolutional neural networks (CNN) have no memory function, so it is difficult to model video frames that are constantly changing in the time domain with CNN.

To this end, dynamic saliency models leverage both static and temporal features to predict human gaze fixation on videos (Gao et al., 2007; Guo and Zhang, 2009; Mahadevan and Vasconcelos, 2009; Rudoy et al., 2013; Seo and Milanfar, 2009; Hou and Zhang, 2008; Fang et al., 2014; Hossein Khatoonabadi et al., 2015; Leboran et al., 2016). Some of these studies (Gao et al., 2007; Mahadevan and Vasconcelos, 2009; Seo and Milanfar, 2009) can be viewed as extensions of existing static saliency models with additional motion features. Conventionally, video saliency models pair bottom-up feature extraction with an ad-hoc motion estimation that can be performed either by means of optical flow or feature tracking. Frame-differencing (Mech and Wollborn, 1997), background subtraction (Tsai and Lai, 2008), optical flow (Horn and Schunck, 1981) and other methods are used to model spatial and motion information. However, these techniques are known for poor performance, especially in complex scene videos.

In contrast, deep video saliency models learn the whole process end-to-end. Some of these saliency models treat spatial and temporal features separately and fuse these features in the last few layers of the DNN architecture in certain way. Other researches simultaneously model the time and space information, directly letting the network simultaneously learn the time and space information and ensure the time and space consistency.

Research works that treat spatial and temporal information separately base on two-stream network architectures (Bak et al., 2017; Zhao and Wu, 2019) that accounts for color images and motion fields separately, or two-layer LSTM with object information (Jiang et al., 2017; Tang et al., 2018)

As one of the first attempts, (Bak et al., 2017) study the use of deep learning for dynamic saliency prediction and propose the so-called spatio-temporal saliency networks. They applied a two-stream (5 layer each) CNN architecture for video saliency prediction. RGB frames and motion maps were fed to the two streams. They have investigated two different fusion strategies, namely element-wise and convolutional fusion strategies, to integrate spatial and temporal information.

(Jiang et al., 2017) concluded that human attention is mainly drawn to objects and their movement. Hence, they propose object-to-motion convolutional neural network (OM-CNN) to learn spatio-temporal features for predicting the intra-frame saliency via exploring the information of both objectness and object motion. Inter-frame saliency is computed by means of a structure-sensitive ConvLSTM architecture.

(Zhao and Wu, 2019) proposes two modules to extract temporal saliency information and spatial information. Moreover, the saliency dynamic information in time is combined with the spatial static saliency estimation model, which directly produces the spatiotemporal saliency inference. A context-aware pyramid feature extraction (CPFE) module is designed for multi-scale high-level feature maps to capture the rich context features. A channel-wise attention (CA) model and a spatial attention (SA) model are respectively applied to the CPFE feature maps and the low-level feature maps, and then fused to detect salient regions. Finally, an edge preservation loss is proposed to get the accurate boundaries of salient regions.

(Tang et al., 2018) used a multiscale spatiotemporal convolutional ConvLSTM network architecture (MSST-ConvLSTM) to combine temporal and spatial information for video saliency detection. This architecture not only retains the original temporal clues but also uses the temporal information in the optical flow map and the structure of LSTM. This part of the study separately learns the information in the time domain and the space domain through neural networks. Generally, to model the information in the time domain, some preprocessing methods, such as the optical flow method, are used. Additionally, the fusion of features extracted in the time and space domains also greatly affect the performance of the network. These works show a better performance and demonstrate the potential advantages in applying neural networks to video saliency problem.

Models that simultaneously model the time and space information directly let the network to concurrently learn the time and space information and ensure the time and space consistency. For instance, in reference (Song et al., 2018), the author first used a pyramid dilated convolution module to extract multiscale spatial features and further extracted spatio-temporal information through a bidirectional convective ConvLSTM structure. Ingeniously, the author used the forward output of the ConvLSTM units as input and directly fed it into the backward ConvLSTM units, which increases the capabilities to extract deeper spatiotemporal features.

In reference (Fan et al., 2019), unlike previous video saliency detection with pixel-level datasets, the author collected a densely annotated dataset that covers different scenes, object categories and motion modes. In (Li et al., 2018), the author proposed a flow-guided recurrent neural encoder (FGRNE) architecture, which uses optical flow networks to estimate motion information per frame in the video and sequential feature evolution encoding in terms of LSTM network units to enhance the temporal coherence modeling of the per-frame feature representation.

(Chaabouni et al., 2016) employed transfer learning to adapt a previously trained deep network for saliency prediction in natural videos. They trained a 5-layer CNN on RGB color planes and residual motion for each video frame. However, their model uses only the very short-term temporal relations of two consecutive frames. In (Bazzani et al., 2016), a recurrent mixture density network is proposed for saliency prediction. The input clip of 16 frames is fed to a 3D CNN, whose output becomes the input to a LSTM. Finally, a linear layer projects the LSTM representation to a Gaussian mixture model, which describes the saliency map. In a similar vein, (Mnih et al., 2014) applied LSTMs to predict video saliency maps, relying on both short- and long-term memory of attention deployment.

In (Leifman et al., 2017), RGB color planes, dense optical flow map, depth map and the previous saliency map are fed to a 7-layered encoder-decoder structure to predict fixations of observers who viewed RGBD videos on a 2D screen.

As in their previous work (Gorji and Clark, 2018), here they used a multi-stream ConvLSTM to augment state-of-the-art static saliency models with dynamic attentional push (shared attention). Their network contains a saliency pathway and three push pathways including gaze following, rapid scene changes, and attentional bounce. The multi-pathway structure is followed by a CNN that learns to combine the complementary and time-varying outputs of the CNN-LSTMs by minimizing the relative entropy between the augmented saliency and viewers fixations on videos.

(Wang et al., 2018b), proposed the Attentive CNN-LSTM Network which augments a CNN-LSTM with a supervised attention mechanism to enable fast end-to-end saliency learning. The attention mechanism explicitly encode static saliency information allowing LSTM to focus on learning a more flexible temporal saliency representation across successive frames. Such a design fully leverages existing large-scale static fixation datasets, avoids overfitting, and significantly improves training efficiency.

(Sun et al., 2018) proposed a robust deep model that utilizes memory and motion information to capture salient points across successive frames. The memory information was exploited to enhance the model generalization by considering the fact that changes between two adjacent frames are limited within a certain range, and hence the corresponding fixations should remain correlated.

There are some more salient object detection models (Liu et al., 2010; Achanta et al., 2009; Cheng et al., 2014; Wang et al., 2015; Wang et al., 2017b; Borji et al., 2015; Hou et al., 2017) that attempt to uniformly highlight salient object regions in images or videos. Those models are often task-driven and focus on inferring the main object, in stead of investigating the behavior of the HVS during scene free viewing.

## 2.2 Video Saliency Dataset

Recent advances in the area of human attention and dynamic fixation prediction are primarily triggered by the release of improved and large saliency dataset (Hadizadeh et al., 2011; Itti, 2004; Mathe and Sminchisescu, 2014; Mital et al., 2011). These dataset improved the understanding of human visual attention and boosted the performance of computational models.

The DHF1K (Wang et al., 2018b) dataset provide human fixations on a more diverse and representative dynamic nature scenes while free-viewing. DHF1K includes 1K video sequences annotated by 17 observers with an eye-tracker device.In DHF1K, each video was manually annotated with a category label, which was further classified into 7 main categories: daily activity, sport, social activity, artistic performance, animal artifact and scenery.

The Hollywood-2 (Mathe and Sminchisescu, 2014) provide a dataset with 12 classes of human actions and 10 classes of scenes distributed over 3669 video clips and approximately 20.1 hours of video in total. The dataset intends to provide a comprehensive benchmark for human action recognition in realistic and challenging settings. According to analysis conducted by (Xie et al., 2020), 84.5 fixations Hollywood-2 dataset are located around the faces.

The UCF Sports dataset (Mathe and Sminchisescu, 2014) consists of a set of actions collected from various sports which are typically featured on broadcast television channels such as the BBC and

ESPN. The video sequences were obtained from a wide range of stock footage websites including BBC Motion gallery and GettyImages. It contains 150 videos taken from the UCF sports action dataset (Rodriguez et al., 2008). According to (Xie et al., 2020), 82.3 fixations of UCF sports saliency dataset fall inside the human body area.

Other datasets are either limited in terms of variety and scale of video stimuli (Mital et al., 2011; Itti, 2004; Hadizadeh et al., 2011; Bylinskii et al., 2015; Huang et al., 2015), or collected for a special purpose (e.g., salient objects in videos (Wang et al., 2015)). More importantly, none of the aforementioned datasets includes a preserved test set for avoiding potential data overfitting, which has seriously hampered the research process.

# 3 OUR APPROACH

## 3.1 Overview

We propose a novel stacked-ConvLSTM based video saliency prediction model. Fig. 1 depicts the architecture of our video saliency prediction model. It is a stacked-ConvLSTM architecture that use both convolutional and recurrent networks. Input to our stacked-ConvLSTM are preprocessed using a novel XY-shift frame differencing layer. This layer takes an absolute difference of an image and its shifted copy and return a high-pass filtered map. Furthermore, a three-frame differencing method takes this data and provide a temporal information aware spatial data map. Three-frame differencing help to magnify the effect of temporal features on the spatial domain and boost the capacity of the stacked-ConvLSTM component on spatio-temporal saliency prediction. Thus, our model produce accurate video saliency prediction with improved generalization. In this section, we introduce our proposed model architecture, and its three important components, namely the stacked-ConvLSTM module, the VGG-16 (Simonyan and Zisserman, 2014), and the XY-shift frame differencing module in detail.

## 3.2 The Stacked-ConvLSTM Model

Fig 1 shows our proposed framework, consisting of three parts: the static convolutional component based on VGG-16 and with the weights of ImageNet (Deng et al., 2009), XY-shift frame differencing and the stacked-ConvLSTM component.

## 3.3 Implementation Details

The implementation details are as follows. First, two-stream of data are passed to the VGG-16 and frame differencing components. The VGG-16 (Simonyan and Zisserman, 2014) extract spatial features from the raw image frames. In order to preserve more spatial details, Pool 4 and Pool 5 layers are removed, resulting in x8 instead of ×32 downsampling. At time step t, the input RGB image $X_t$ size is (224×224×3). The output characteristic size of this component is [32, 40, 512]. Concurrently, we apply a batch level XY-shift frame differencing and three-frame differencing on each members of a batch to magnify temporal features on spatial domain. The XY-shift frame differencing differs a frame from its shifted replica. The effect of this operation is equivalent to the result of a high-pass filter method, but with significantly smaller computational resource. We have mainly used this method to reduce the visibility of irrelevant background objects and expose foreground objects. The mathematical formalization of XY-shift frame differencing is depicted as follows in equation 1. Let $a$ be the first channel of image $A$ with a shape of (h,w,3). Then, the XY-shift frame differencing of $a$ is calculated as:

$$g(a) = \begin{cases} a(x_i, y_j) - a(x_{i+f}+, y_{j+f}), & \text{if} \\ i <= h - f and j <= w - f & \\ a(x_i, y_j) - a(x_{i-f}+, y_{j-f}), & \text{if i = h or j = w.} \end{cases} \tag{1}$$

where h and w stands for the height and width of the channel and f is a shift factor.

What follows the XY-shift frame differencing is an improved three-frame differencing technique. This technique use the output of XY-shift differencing. It takes three consecutive frames, compute the difference between the current frame and the previous frame, the current frame and the next frame separately, and extract a pixel-wise max between these two resulting frames. This technique is adapted and enhanced to improve the extraction of temporal features from datasets in spatio-temporal domain. The improved three-frame differencing method is formalized as follows in equation 2. Let A,B, and C be the first channel of three consecutive XY-shift frame differenced frames with a shape of (h,w). Let B be the first channel of the current frame. Then the improved three-frame differencing, f(A,B,C), is calculated as:

$$f(A,B,C)_{i,j} = max_{i,j}(|B_{i,j} - A_{i,j}|, |B_{i,j} - C_{i,j}|) \tag{2}$$

where for i,j >= 0 and i<= h and j <= w.

Furthermore, the pixel-wise maximum of two images is computed as shown in 3. Let Q1 be the absolute difference of the current frame B and its predecessor frame A. Let Q2 be the absolute difference

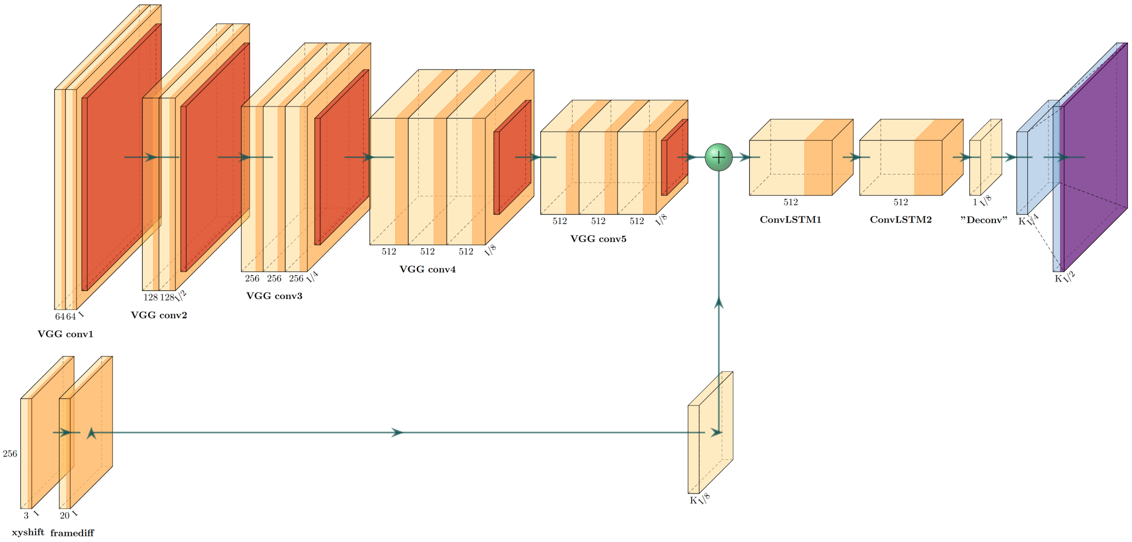

Figure 1: Interactive Video Saliency Identification With Attentive ConvLSTM Architecture.

of the current frame B and its successor frame C. Let's say both differenced images have a size of (h,w). Then, the pixel-wise maximum, $P_max$, of these two frames is calculated as:

$$max(Q1,Q2)_{i,j} = \begin{cases} Q1_{i,j}, & \text{if } Q_{i,j} > Q2_{i,j} \\ Q2_{i,j}, & \text{if otherwise} \end{cases} \quad (3)$$

where for i,j >= 0 and i<= h and j <= w.

A residual layer fusing the VGG-16 extracted spatial features and frame differencing output frames is applied succeeding the aforementioned components. Finally, the output of both VGG-16 and frame differencing mixed layer is deep fused into a single feature space. A [30x40x512] output of the residual layer is further fed to our stacked-ConvLSTM network. The main reason for stacking ConvLSTM is to allow for greater model complexity. Even though there are large-scale datasets like DHF1K that have 1K videos, the amount of training data is still insufficient, considering the high correlation among frames within same video (Jiang et al., 2017). Hence, increasing the complexity of the model help to extract more complex features in return providing robust video saliency prediction model. The size of the feature map after the stacked-ConvLSTM is 32x40x256. By passing this output through a convolutional layer, with kernel size 1x1, and upsampling the resulting feature map, we get 128x160x1 and 64x80x1 saliency map corresponding to the different loss functions we employed in this research work.

## 3.4 Loss Functions

To better generate robust saliency maps, we use three loss functions as used in (Jiang et al., 2018) and (Wang et al., 2018b). Linear Correlation Coefficient(CC) (Jost et al., 2005), the Kullback-Leibler divergence (KLD) (Tatler et al., 2005) and Normalized Scanpath Saliency (NSS) (Peters et al., 2005). The essence of using multiple loss functions is to increase the degree of learning and generalization of the model.

We denote the predicted saliency map as $Y \in [0,1]^{28x28}$, the map of fixation locations as $P \in \{0,1\}^{28x28}$ and the continuous saliency map (distribution) as $Q \in [0,1]^{28x28}$. Here the fixation map P is discrete, that records whether a pixel receives human fixation. The continuous saliency map is obtained via blurring each fixation location with a small Gaussian kernel. Our loss functions is defined as follows:

$$L(Y,P,Q) = L_{KL}(Y,Q) + \alpha_1 L_{CC}(Y,Q) + \alpha_2 L_{NSS}(Y,P) \quad (4)$$

where $L_{KL}, L)CCandL_{NSS}$ are the Kullback-Leibler (KL) divergence, the Linear Correlation Coefficient (CC), and the Normalized Scanpath Saliency (NSS), respectively, which are derived from commonly used metrics to evaluate saliency prediction models. $\alpha s$ are balance parameters and are empirically set to $\alpha_1 = \alpha_2 = 0.1$.

Kullback–Leibler divergence (KLD) measures the divergence between the distribution S and $\hat{S}$:

$$L_{KL}(S,\hat{S}) = \sum_{i=1}^{NXM} \hat{S}_i \log \frac{\hat{S}_i}{Si} \quad (5)$$

Normalized Scanpath Saliency metric was introduced in (Peters et al., 2005), to evaluate the degree of congruency between human eye fixations and a predicted saliency map. Instead of relying on a saliency map as ground truth, the predictions are evaluated against the true fixations map. The value of the saliency map at each fixation point is normalized with the whole saliency map variance:

$$L_{NSS}(S^{fix},\hat{S}) = \frac{1}{NXM} \sum_{i=1}^{NXM} [\frac{\hat{S}_i - \mu(\hat{S}_i)}{\alpha(\hat{S}_i)}] S_i^{fix} \quad (6)$$

Pearson's Correlation Coefficient (CC) measures the linear correlation between the ground truth saliency map and the predicted saliency map:

$$L_{CC}(S,\hat{S}) = \frac{\alpha(S,\hat{S})}{\alpha(S)\alpha(\hat{S})} \quad (7)$$

## 3.5 Training Protocol

Our model is iteratively trained with sequential fixation and image data. In training, a video training batch is cascaded with an image training batch. More specifically, in a video training batch, we apply a loss defined over the final dynamic saliency prediction from LSTM. For each video training batch, 20 consecutive frames from the same video are used. Both the video and the start frames are randomly selected. For each image training batch, we set the batch size as 20, and the images are randomly sampled from existing static fixation dataset.

# 4 EXPERIMENTS

## 4.1 Datasets and Evaluation Mertrics

### 4.1.1 Datasets

We use the DHF1K (Wang et al., 2018b) dataset for training and evaluation. We use only the first 70% of the DHF1K dataset and used 70%/10%/30% training/validation/testing ratio to split data for the experiment. Hence, our model is trained and validated on 420 and 70 randomly selected videos. Moreover, the evaluation of our proposed model is undertaken on 210 test video sequences.

### 4.1.2 Evaluation Metrics

We use five performance evaluation metrics, namely Normalized Scanpath Saliency (NSS), Similarity Metric (SIM), Linear Correlation Coefficient (CC), AUC-Judd (AUC-J), and shuffled AUC (s-AUC).

### 4.1.3 Competitors

To prove the effectiveness of our proposed model, we compare our model with sixteen saliency models. Among them, (Wang et al., 2018b), PQFT (Guo and Zhang, 2009), Seo et al. (Seo and Milanfar, 2009), Rudoy et al.(Rudoy et al., 2013), Hou et al. (Hou and Zhang, 2008), Fang et al. (Fang et al., 2014), OBDL (Hossein Khatoonabadi et al., 2015), AWS-D (Leboran et al., 2016), OM-CNN (Jiang et al., 2017), and Two-stream (Bak et al., 2017) are dynamic saliency models. Furthermore, ITTI (Itti et al., 1998), GBVS (Harel et al., 2006), SALICON (Huang et al., 2015), DVA (Wang and Shen, 2017), Shallow-Net (Pan et al., 2016), and Deep-Net (Pan et al., 2016) are state-of-the-art static attention models. OM-CNN, Two-stream, SALICON, DVA, Shallow-Net, and Deep-Net are deep learning models, and others are classical saliency models. We choose these models due to publicly available implementations and their representability of the state-of-the-art.

### 4.1.4 Computational Load

The whole model is trained in an end-to-end manner. The entire training procedure takes about 60 hours with a single NVIDIA Quadro RTX 3000 Max-Q GPU. Our model takes about 0.84s to process a frame image of size 224 × 224.

## 4.2 Performance Comparison

### 4.2.1 Performance on DHF1K

Table 1 presents the comparative performance of our model against the competitor models. It is observed that our model significantly outperformed all static saliency models and the majority of dynamic models, across all performance metrics. Our model show competitive result with the one reported in (Wang et al., 2018b). This is directly attributed to the novel XY-shift frame differencing technique and stacked-ConvLSTM network incorporated in our architecture.

## 4.3 Analysis

In the course of our research, we have conducted extensive experiments. Here, we analyse our model and

Table 1: Quantitative results on DHF1K: Training setting I is trained and evaluated using only DHF1K dataset.

| | Models/Datasets | DHF1K | | | | |
|---|---|---|---|---|---|---|
| | | AUC-J | SIM | s-AUC | CC | NSS |
| Dynamic models | (Guo and Zhang, 2009) | 0.699 | 0.139 | 0.562 | 0.137 | 0.749 |
| | (Seo and Milanfar, 2009) | 0.635 | 0.142 | 0.499 | 0.070 | 0.334 |
| | (Rudoy et al., 2013) | 0.769 | 0.214 | 0.501 | 0.285 | 1.498 |
| | (Hou and Zhang, 2008) | 0.726 | 0.167 | 0.545 | 0.150 | 0.847 |
| | (Fang et al., 2014) | 0.819 | 0.198 | 0.537 | 0.273 | 1.539 |
| | (Hossein Khatoonabadi et al., 2015) | 0.638 | 0.171 | 0.500 | 0.117 | 0.495 |
| | (Leboran et al., 2016) | 0.703 | 0.157 | 0.513 | 0.174 | 0.940 |
| | (Jiang et al., 2017) | 0.856 | 0.256 | 0.583 | 0.344 | 1.911 |
| | (Bak et al., 2017) | 0.834 | 0.197 | 0.581 | 0.325 | 1.632 |
| | (Wang et al., 2018b) | **0.885** | **0.311** | 0.553 | **0.415** | **2.259** |
| Static models | (Itti et al., 1998) | 0.774 | 0.162 | 0.553 | 0.233 | 1.207 |
| | (Harel et al., 2006) | 0.828 | 0.186 | 0.554 | 0.283 | 1.474 |
| | (Huang et al., 2015) | 0.857 | 0.232 | 0.590 | 0.327 | 1.901 |
| | (Pan et al., 2016) Shallow-Net | 0.833 | 0.182 | 0.529 | 0.295 | 1.509 |
| | (Pan et al., 2016) Deep-Net | 0.855 | 0.201 | 0.592 | 0.331 | 1.775 |
| | (Wang and Shen, 2017) | 0.860 | 0.262 | 0.595 | 0.358 | 2.013 |
| Training Setting I | Our model | **0.878** | **0.304** | **0.665** | **0.405** | **2.239** |

competitive models thoroughly with the intention of giving deeper insight to the state-of-the-art models and suggest opportunities that we believe are inspiring for future work in dynamic video prediction.

We conduct our analysis first by contrasting the effect of employing deep learning methods for static and dynamic saliency prediction. According to our finding, deep learning methods outperform classical methods both in static DVA (Wang and Shen, 2017), Deep-Net (Pan et al., 2016) and dynamic OM-CNN (Jiang et al., 2017), Two-stream (Bak et al., 2017), ACL (Wang et al., 2018b) saliency prediction problems, and in almost all saliency prediction metrics. On the other hand, classical methods show relatively reduced performance in static saliency predication ITTI (Itti et al., 1998),GBVS (Harel et al., 2006). A significant performance degradation is observed when static saliency prediction algorithms are employed for dynamic saliency prediction problem sets PQFT (Guo and Zhang, 2009), (Seo and Milanfar, 2009), (Rudoy et al., 2013), (Hou and Zhang, 2008), (Fang et al., 2014). This demonstrates the strong learning ability of deep neural network and the promise of developing deep learning network based models in this challenging area. Moreover, the analyses show the inherent incapability of classic machine learning methods for complex problem sets such as, saliency prediction.

## 4.4 Ablation Study

In this section, we discuss component wise contribution of our model. We verify the effectiveness of various components and their order of composition in our model.

The effectiveness of the XY-shift frame differencing technique is analyzed by eliminating its effect from the general architecture. A stacked-ConvLSTM architecture without our novel frame differencing layer show reduced performance in capturing saliency in highly dynamic scenes. Quantitatively speaking, we noticed 20 to 25 percent performance reduction in all evaluation metrics we employed. Performance gains due to the novel XY-shift frame differencing is attributed to the magnified temporal features in the spatial domain. Magnifying temporal features in the spatial domain help the stacked-ConvLSTM component to easily extract spatio-temporal saliency features.

Besides, due to the complex nature of dynamic saliency prediction, the use of stacked-ConvLSTM component right after a spatial feature extractor component improve our model's performance on complex feature extraction. Consequently, the use of stacked-ConvLSTM rather than a single ConvLSTM architecture show slight performance improvement.

Another interesting finding in the course of our research is the effect of residual layer positioning. The variation in the position of residual layers show significant performance variation. We placed residual layers residual layers in different positions, such as at the end of the primary convolutional base, between the ConvLSTM layer, and finally, at the end of our overall encoder, processing every input in a separate stream. Placing residual layer at the beginning of the

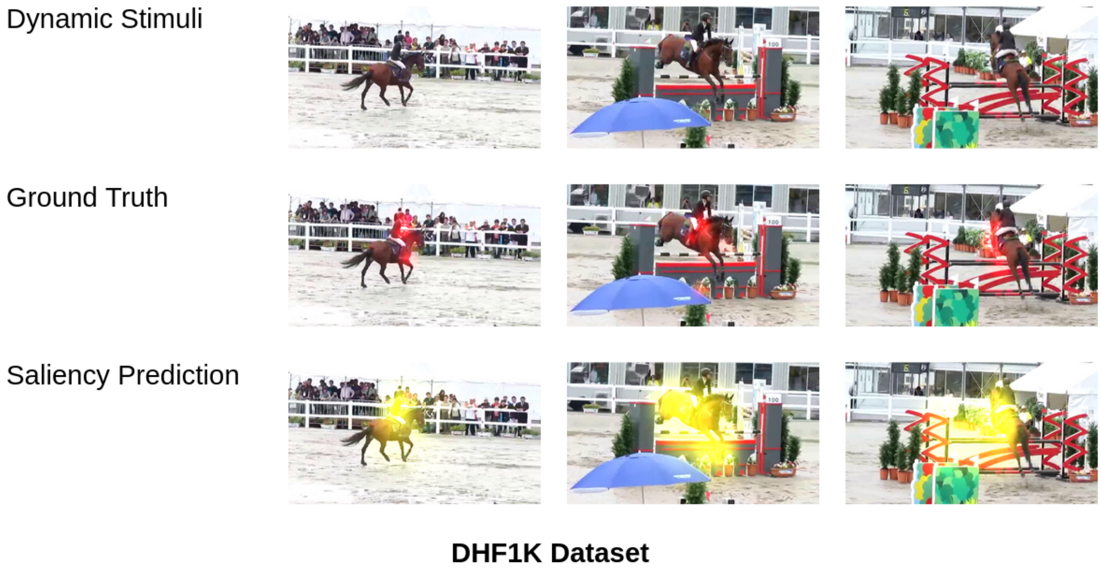

Figure 2: Qualitative results of our video saliency model on DHF1K Dataset.

stacked-ConvLSTM component yield better saliency prediction performance and relatively better resource utilization.

Similarly, we undertook a through qualitative analysis by randomly selecting sequence of frames from our testing set. On the other hand, the interactivity (Wondimu et al., 2022) of our model is evaluated by deploying it in a resource constrained robot called Pepper. The results show the effectiveness of our video saliency prediction model relative to the state-of-the-art video saliency prediction models. Moreover,

## 5 CONCLUSION

In this research, we proposed a novel deep learning based dynamic saliency prediction model, which employ the benefits of a novel XY-shift frame differencing technique and stacked-ConvLSTM network. An extensive experimentation on the largest video saliency dataset, DHF1K (Wang et al., 2018b) is undertaken. We compared our results with similar deep learning based dynamic saliency models. Our experimental results show the effectiveness and superiority of our model against 15 state-of-the-art models and its competitiveness against the outperforming dynamic saliency prediction model (Wang et al., 2018b).

## ACKNOWLEDGEMENTS

This work would not have been possible without the financial support of the French Embassy in Ethiopia, Brittany region administration and the Ethiopia Ministry of Education (MoE). We are also indebted to Brest National School of Engineering (ENIB) and specifically LAB-STICC for creating such a conducive research environment.

## REFERENCES

Achanta, R., Hemami, S., Estrada, F., and Susstrunk, S. (2009). Frequency-tuned salient region detection. In *2009 IEEE conference on computer vision and pattern recognition*, pages 1597–1604. IEEE.

Amso, D. and Scerif, G. (2015). The attentive brain: insights from developmental cognitive neuroscience. *Nature Reviews Neuroscience*, 16(10):606–619.

Bak, C., Kocak, A., Erdem, E., and Erdem, A. (2017). Spatio-temporal saliency networks for dynamic saliency prediction. *IEEE Transactions on Multimedia*, 20(7):1688–1698.

Bazzani, L., Larochelle, H., and Torresani, L. (2016). Recurrent mixture density network for spatiotemporal visual attention. *arXiv preprint arXiv:1603.08199*.

Bohic, M. and Abraira, V. E. (2022). Wired for social touch: the sense that binds us to others. *Current Opinion in Behavioral Sciences*, 43:207–215.

Borji, A., Cheng, M.-M., Jiang, H., and Li, J. (2015). Salient object detection: A benchmark. *IEEE transactions on image processing*, 24(12):5706–5722.

Borji, A. and Itti, L. (2012). State-of-the-art in visual attention modeling. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):185–207.

Bruce, N. and Tsotsos, J. (2005). Saliency based on information maximization. *Advances in neural information processing systems*, 18.

Butko, N. J., Zhang, L., Cottrell, G. W., and Movellan, J. R. (2008). Visual saliency model for robot cameras. In *2008 IEEE International Conference on Robotics and Automation*, pages 2398–2403. IEEE.

Bylinskii, Z., Judd, T., Borji, A., Itti, L., Durand, F., Oliva, A., and Torralba, A. (2015). Mit saliency benchmark. *MIT Press*.

Chaabouni, S., Benois-Pineau, J., and Amar, C. B. (2016). Transfer learning with deep networks for saliency prediction in natural video. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 1604–1608. IEEE.

Chang, C.-K., Siagian, C., and Itti, L. (2010). Mobile robot vision navigation & localization using gist and saliency. In *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4147–4154. IEEE.

Chen, Y., Zhang, W., Wang, S., Li, L., and Huang, Q. (2018). Saliency-based spatiotemporal attention for video captioning. In *2018 IEEE fourth international conference on multimedia big data (BigMM)*, pages 1–8. IEEE.

Cheng, M.-M., Mitra, N. J., Huang, X., Torr, P. H., and Hu, S.-M. (2014). Global contrast based salient region detection. *IEEE transactions on pattern analysis and machine intelligence*, 37(3):569–582.

Cornia, M., Baraldi, L., Serra, G., and Cucchiara, R. (2018). Paying more attention to saliency: Image captioning with saliency and context attention. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 14(2):1–21.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Diaz, C. G., Perry, P., and Fiebrink, R. (2019). Interactive machine learning for more expressive game interactions. In *2019 IEEE Conference on Games (CoG)*, pages 1–2. IEEE.

Fan, D.-P., Wang, W., Cheng, M.-M., and Shen, J. (2019). Shifting more attention to video salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8554–8564.

Fang, Y., Wang, Z., Lin, W., and Fang, Z. (2014). Video saliency incorporating spatiotemporal cues and uncertainty weighting. *IEEE transactions on image processing*, 23(9):3910–3921.

Ferreira, J. F. and Dias, J. (2014). Attentional mechanisms for socially interactive robots–a survey. *IEEE Transactions on Autonomous Mental Development*, 6(2):110–125.

Fukuchi, K., Miyazato, K., Kimura, A., Takagi, S., and Yamato, J. (2009). Saliency-based video segmentation with graph cuts and sequentially updated priors. In *2009 IEEE International Conference on Multimedia and Expo*, pages 638–641. IEEE.

Gao, D., Mahadevan, V., and Vasconcelos, N. (2007). The discriminant center-surround hypothesis for bottom-up saliency. *Advances in neural information processing systems*, 20.

Gorji, S. and Clark, J. J. (2018). Going from image to video saliency: Augmenting image salience with dynamic attentional push. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7501–7511.

Guo, C. and Zhang, L. (2009). A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. *IEEE transactions on image processing*, 19(1):185–198.

Hadizadeh, H., Enriquez, M. J., and Bajic, I. V. (2011). Eye-tracking database for a set of standard video sequences. *IEEE Transactions on Image Processing*, 21(2):898–903.

Harel, J., Koch, C., and Perona, P. (2006). Graph-based visual saliency. *Advances in neural information processing systems*, 19.

Horn, B. K. and Schunck, B. G. (1981). Determining optical flow. *Artificial intelligence*, 17(1-3):185–203.

Hossein Khatoonabadi, S., Vasconcelos, N., Bajic, I. V., and Shan, Y. (2015). How many bits does it take for a stimulus to be salient? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5501–5510.

Hou, Q., Cheng, M.-M., Hu, X., Borji, A., Tu, Z., and Torr, P. H. (2017). Deeply supervised salient object detection with short connections. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3203–3212.

Hou, X. and Zhang, L. (2008). Dynamic visual attention: Searching for coding length increments. *Advances in neural information processing systems*, 21.

Huang, X., Shen, C., Boix, X., and Zhao, Q. (2015). Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *Proceedings of the IEEE international conference on computer vision*, pages 262–270.

Itti, L. (2004). Automatic foveation for video compression using a neurobiological model of visual attention. *IEEE transactions on image processing*, 13(10):1304–1318.

Itti, L. and Koch, C. (2001). Computational modelling of visual attention. *Nature reviews neuroscience*, 2(3):194–203.

Itti, L., Koch, C., and Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 20(11):1254–1259.

Ji, J., Xiang, K., and Wang, X. (2022). Scvs: blind image quality assessment based on spatial correlation and visual saliency. *The Visual Computer*, pages 1–16.

Jiang, H., Wang, J., Yuan, Z., Wu, Y., Zheng, N., and Li, S. (2013). Salient object detection: A discriminative regional feature integration approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2083–2090.

Jiang, L., Xu, M., Liu, T., Qiao, M., and Wang, Z. (2018). Deepvs: A deep learning based video saliency prediction approach. In *Proceedings of the european conference on computer vision (eccv)*, pages 602–617.

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

Jost, T., Ouerhani, N., Von Wartburg, R., Müri, R., and Hügli, H. (2005). Assessing the contribution of color in visual attention. *Computer Vision and Image Understanding*, 100(1-2):107–123.

Judd, T., Ehinger, K., Durand, F., and Torralba, A. (2009). Learning to predict where humans look. In *2009 IEEE 12th international conference on computer vision*, pages 2106–2113. IEEE.

Kruthiventi, S. S., Ayush, K., and Babu, R. V. (2017). Deepfix: A fully convolutional neural network for predicting human eye fixations. *IEEE Transactions on Image Processing*, 26(9):4446–4456.

Lateef, F., Kas, M., and Ruichek, Y. (2021). Saliency heatmap as visual attention for autonomous driving using generative adversarial network (gan). *IEEE Transactions on Intelligent Transportation Systems*.

Le Meur, O., Le Callet, P., Barba, D., and Thoreau, D. (2006). A coherent computational approach to model bottom-up visual attention. *IEEE transactions on pattern analysis and machine intelligence*, 28(5):802–817.

Leboran, V., Garcia-Diaz, A., Fdez-Vidal, X. R., and Pardo, X. M. (2016). Dynamic whitening saliency. *IEEE transactions on pattern analysis and machine intelligence*, 39(5):893–907.

Leifman, G., Rudoy, D., Swedish, T., Bayro-Corrochano, E., and Raskar, R. (2017). Learning gaze transitions from depth to improve video saliency estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1698–1707.

Li, G., Xie, Y., Wei, T., Wang, K., and Lin, L. (2018). Flow guided recurrent neural encoder for video salient object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3243–3252.

Liu, N., Han, J., Liu, T., and Li, X. (2016). Learning to predict eye fixations via multiresolution convolutional neural networks. *IEEE transactions on neural networks and learning systems*, 29(2):392–404.

Liu, T., Yuan, Z., Sun, J., Wang, J., Zheng, N., Tang, X., and Shum, H.-Y. (2010). Learning to detect a salient object. *IEEE Transactions on Pattern analysis and machine intelligence*, 33(2):353–367.

Mahadevan, V. and Vasconcelos, N. (2009). Spatiotemporal saliency in dynamic scenes. *IEEE transactions on pattern analysis and machine intelligence*, 32(1):171–177.

Marat, S., Ho Phuoc, T., Granjon, L., Guyader, N., Pellerin, D., and Guérin-Dugué, A. (2009). Modelling spatio-temporal saliency to predict gaze direction for short videos. *International journal of computer vision*, 82(3):231–243.

Mathe, S. and Sminchisescu, C. (2014). Actions in the eye: Dynamic gaze datasets and learnt saliency models for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(7):1408–1424.

Mech, R. and Wollborn, M. (1997). A noise robust method for segmentation of moving objects in video sequences. In *1997 IEEE International conference on acoustics, speech, and signal processing*, volume 4, pages 2657–2660. IEEE.

Mital, P. K., Smith, T. J., Hill, R. L., and Henderson, J. M. (2011). Clustering of gaze during dynamic scene viewing is predicted by motion. *Cognitive computation*, 3(1):5–24.

Mnih, V., Heess, N., Graves, A., et al. (2014). Recurrent models of visual attention. *Advances in neural information processing systems*, 27.

Pal, A., Mondal, S., and Christensen, H. I. (2020). " looking at the right stuff"-guided semantic-gaze for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11883–11892.

Pan, J., Sayrol, E., Giro-i Nieto, X., McGuinness, K., and O'Connor, N. E. (2016). Shallow and deep convolutional networks for saliency prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 598–606.

Peters, R. J., Iyer, A., Itti, L., and Koch, C. (2005). Components of bottom-up gaze allocation in natural images. *Vision research*, 45(18):2397–2416.

Rice, L., Wong, E., and Kolter, Z. (2020). Overfitting in adversarially robust deep learning. In *International Conference on Machine Learning*, pages 8093–8104. PMLR.

Roberts, R., Ta, D.-N., Straub, J., Ok, K., and Dellaert, F. (2012). Saliency detection and model-based tracking: a two part vision system for small robot navigation in forested environment. In *Unmanned Systems Technology XIV*, volume 8387, page 83870S. International Society for Optics and Photonics.

Rodriguez, M. D., Ahmed, J., and Shah, M. (2008). Action mach a spatio-temporal maximum average correlation height filter for action recognition. In *2008 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE.

Rudoy, D., Goldman, D. B., Shechtman, E., and Zelnik-Manor, L. (2013). Learning video saliency from human gaze using candidate selection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1147–1154.

Schauerte, B. and Stiefelhagen, R. (2014). "look at this!" learning to guide visual saliency in human-robot interaction. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 995–1002. IEEE.

Schillaci, G., Bodiroža, S., and Hafner, V. V. (2013). Evaluating the effect of saliency detection and attention manipulation in human-robot interaction. *International Journal of Social Robotics*, 5(1):139–152.

Seo, H. J. and Milanfar, P. (2009). Static and space-time visual saliency detection by self-resemblance. *Journal of vision*, 9(12):15–15.

Shao, Z., Wang, L., Wang, Z., Du, W., and Wu, W. (2019). Saliency-aware convolution neural network for ship detection in surveillance video. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(3):781–794.

Shi, J., Yan, Q., Xu, L., and Jia, J. (2015a). Hierarchical image saliency detection on extended cssd. *IEEE transactions on pattern analysis and machine intelligence*, 38(4):717–729.

Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., and Woo, W.-c. (2015b). Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Song, H., Wang, W., Zhao, S., Shen, J., and Lam, K.-M. (2018). Pyramid dilated deeper convlstm for video salient object detection. In *Proceedings of the European conference on computer vision (ECCV)*, pages 715–731.

Sun, M., Zhou, Z., Hu, Q., Wang, Z., and Jiang, J. (2018). Sg-fcn: A motion and memory-based deep learning model for video saliency detection. *IEEE transactions on cybernetics*, 49(8):2900–2911.

Tang, Y., Zou, W., Jin, Z., and Li, X. (2018). Multi-scale spatiotemporal conv-lstm network for video saliency detection. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, pages 362–369.

Tatler, B. W., Baddeley, R. J., and Gilchrist, I. D. (2005). Visual correlates of fixation selection: Effects of scale and time. *Vision research*, 45(5):643–659.

Tsai, D.-M. and Lai, S.-C. (2008). Independent component analysis-based background subtraction for indoor surveillance. *IEEE Transactions on image processing*, 18(1):158–167.

Vig, E., Dorr, M., and Cox, D. (2014). Large-scale optimization of hierarchical features for saliency prediction in natural images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2798–2805.

Wang, H., Xu, Y., and Han, Y. (2018a). Spotting and aggregating salient regions for video captioning. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1519–1526.

Wang, T., Borji, A., Zhang, L., Zhang, P., and Lu, H. (2017a). A stagewise refinement model for detecting salient objects in images. In *Proceedings of the IEEE international conference on computer vision*, pages 4019–4028.

Wang, W. and Shen, J. (2017). Deep visual attention prediction. *IEEE Transactions on Image Processing*, 27(5):2368–2378.

Wang, W., Shen, J., Guo, F., Cheng, M.-M., and Borji, A. (2018b). Revisiting video saliency: A large-scale benchmark and a new model. In *Proceedings of the IEEE Conference on computer vision and pattern recognition*, pages 4894–4903.

Wang, W., Shen, J., and Porikli, F. (2015). Saliency-aware geodesic video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3395–3402.

Wang, W., Shen, J., and Shao, L. (2017b). Video salient object detection via fully convolutional networks. *IEEE Transactions on Image Processing*, 27(1):38–49.

Wang, W., Shen, J., Yu, Y., and Ma, K.-L. (2016). Stereoscopic thumbnail creation via efficient stereo saliency detection. *IEEE transactions on visualization and computer graphics*, 23(8):2014–2027.

Wondimu, N. A., Buche, C., and Visser, U. (2022). Interactive machine learning: A state of the art review. *arXiv preprint arXiv:2207.06196*.

Xie, J., Cheng, M.-M., Ling, H., and Borji, A. (2020). Revisiting video saliency prediction in the deep learning era. *IEEE transactions on pattern analysis and machine intelligence*.

Xie, Y. and Lu, H. (2011). Visual saliency detection based on bayesian model. In *2011 18th IEEE International Conference on Image Processing*, pages 645–648. IEEE.

Yang, C., Zhang, L., Lu, H., Ruan, X., and Yang, M.-H. (2013). Saliency detection via graph-based manifold ranking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3166–3173.

Yubing, T., Cheikh, F. A., Guraya, F. F. E., Konik, H., and Trémeau, A. (2011). A spatiotemporal saliency model for video surveillance. *Cognitive Computation*, 3(1):241–263.

Yun, I., Jung, C., Wang, X., Hero, A. O., and Kim, J. K. (2019). Part-level convolutional neural networks for pedestrian detection using saliency and boundary box alignment. *IEEE Access*, 7:23027–23037.

Zhang, P., Wang, D., Lu, H., Wang, H., and Ruan, X. (2017). Amulet: Aggregating multi-level convolutional features for salient object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 202–211.

Zhang, P., Zhuo, T., Huang, H., and Kankanhalli, M. (2018). Saliency flow based video segmentation via motion guided contour refinement. *Signal Processing*, 142:431–440.

Zhao, T. and Wu, X. (2019). Pyramid feature attention network for saliency detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3085–3094.