Enhanced Human-Robot Interaction through Spatio-Temporal Saliency Prediction

Natnael WONDIMU[1,4*], Ubbo VISSER[3*†] and Cédric BUCHE[1,2*†]

[1*]Lab-STICC, Brest National School of Engineering, 945 Av. du Technopôle, Plouzané, 29280, Brittany, France.
[2]IRL CROSSING, CNRS, University of Adelaide, Brailsford Robertson Building, Adelaide, 5000, SA, Australia.
[3]Computer Science, University of Miami, 1320 S Dixie Hwy, Coral Gables, 33124, Florida, USA.
[4]School of Information Technology and Engineering, Addis Ababa University, NBH1, 4killo King George VI St, Addis Ababa, 1000, Addis Ababa, Ethiopia.

*Corresponding author(s). E-mail(s): natnael.argaw@aau.edu.et; visser@cs.miami.edu; buche@enib.fr;
†These authors contributed equally to this work.

Abstract

Spatio-temporal saliency prediction, with its broad applications ranging from human-robot interaction to social standard robotics, plays a significant role in computer vision. Traditional static saliency prediction techniques are inadequate for video saliency prediction, necessitating the utilization of deep learning approaches that leverage spatio-temporal saliency. However, deep learning techniques encounter challenges due to the complex nature of video saliency prediction and the scarcity of representative saliency benchmarks and architectures designed specifically for video saliency. In this paper, we propose a novel video saliency prediction model that incorporates a stacked-ConvLSTM-based architecture along with a custom XY-shift frame differencing layer. Our model adopts an encoder-decoder structure, incorporating a prior layer for XY-shift frame differencing, a residual layer that combines spatially processed features based on the VGG-16 model with XY-shift frame-differenced frames, and a stacked-ConvLSTM component. Furthermore, we apply our video saliency prediction model to enhance human-robot interaction. We conducted comprehensive evaluations and comparisons with state-of-the-art static and dynamic video saliency prediction models, demonstrating competitive results. Additionally, subjective ratings and latency metrics were employed to evaluate the performance of our model in human-robot interaction scenarios using the Pepper Robot, both in virtual and physical environments. The obtained promising results support the efficacy of enhancing

human-robot interaction through saliency prediction models. Finally, we suggest potential research areas that may be of interest to researchers in this field. Our work contributes to the advancement of video saliency prediction and highlights its importance in enhancing human-robot interaction.

Keywords: Spatio-temporal Saliency Prediction, Video Saliency Prediction, Gaze Prediction, Video Saliency Dataset, Human Robot Interaction, Intuitive HRI, Efficient HRI

1 Introduction

In the realm of robotics, it is imperative to develop robust computational models that accurately mimic human perceptual and action intelligence in real-time. Saliency prediction, a fundamental capability of the human visual system, allows for the rapid identification of important scenes within the visual field [1, 2]. By developing computational models that replicate this feature, it becomes possible to enable efficient and realistic human-robot interaction in social standard robotic environments [1, 3, 4]. This ability is crucial for establishing seamless communication and collaboration between humans and robots. By accurately predicting saliency, robots can optimize their perception, decision-making, and action processes, leading to more effective and intelligent behavior. Moreover, saliency prediction models have proven to be valuable resources for enhancing various computer vision tasks, such as video segmentation [5, 6], video captioning [7, 8], autonomous driving [9, 10], and surveillance [11, 12] and other areas [13–17]. Leveraging these models have the potential to revolutionize human-robot interaction in various industrial and social contexts.

Visual saliency research has explored spatial [18, 19] and spatio-temporal [20] perspectives. Early static image saliency prediction models utilized spatial information from individual images or frames, with computational models, particularly those inspired by deep neural networks (DNN), proving effective for static saliency prediction [21–25]. However, when applied to video stimuli, static image saliency prediction models suffer from reduced performance due to the dynamic and spatio-temporal nature of videos.

Recent advancements in video saliency prediction models have addressed this limitation by considering the spatio-temporal aspects of video saliency datasets. This approach is supported by cognitive and neuroscience research, which highlights the significance of spatio-temporal features in data collected across space and time [26, 27]. Furthermore, the progress in deep neural networks has greatly contributed to the development of dynamic saliency prediction models that efficiently handle spatio-temporal data. These DNN-inspired models leverage the capabilities of deep neural networks to effectively analyze and predict saliency in dynamic video sequences. By incorporating spatio-temporal information, these models achieve superior performance compared to their static counterparts, thus advancing the field of video saliency prediction.

However, a significant limitation of these models is their reliance on datasets that lack generic, representative, and diverse instances in unconstrained task-independent scenarios. This lack of diversity and representativeness poses several challenges to the performance and applicability of these models. Firstly, the models are prone to overfitting, a phenomenon where the model becomes overly specialized to the specific characteristics of the training

dataset, resulting in poor generalization to new and unseen data [28]. Overfitting can lead to inflated performance metrics during training but diminished performance when applied to real-world scenarios. To overcome these limitations, it is crucial to develop and utilize video saliency datasets that incorporate a broader range of instances, encompassing diverse scenes, objects, and activities in unconstrained settings [29, 30]. Such datasets would provide a more comprehensive and realistic representation of the challenges encountered in real-world scenarios. By training video saliency models on these diverse datasets, we can enhance their robustness, generalization ability, and applicability to a wider range of practical applications in computer vision and human-robot interaction.

In order to address the challenges in video saliency prediction and extend the application of our resulting model, we propose a novel video saliency prediction model based on a stacked-ConvLSTM architecture. Our model introduces a new custom layer called XY-Shift frame differencing, which enhances the extraction of temporal features within the spatial domain, resulting in improved video saliency prediction. Additionally, we present a novel approach to fuse temporally magnified spatio-temporal features with spatially engineered features obtained from popular spatial feature extractors like VGG-16 [31]. By integrating these two types of features, our model achieves more accurate and robust saliency predictions.

To further evaluate the performance and applicability of our model, we extended its application to the domain of Human-Robot Interaction (HRI). In the HRI setting, we assess the results of applying our saliency model using various metrics, including subjective ratings and latency measurements. These metrics provide valuable insights into the model's effectiveness in enabling intuitive and efficient human-robot interaction by continuously directing the robot's attention to salient regions in its visual field. The evaluation in the HRI context allows us to validate the practical utility and real-time performance of our model.

Through extensive experimentation on the largest video saliency dataset available, DHF1K [30], we demonstrate the competitive performance of our model against state-of-the-art methods in both saliency prediction and its application in the HRI setting. The results obtained from the DHF1K dataset validate the effectiveness and versatility of our proposed architecture, which combines the XY-Shift frame differencing layer, the fusion of spatio-temporal features, and the stacked-ConvLSTM component.

The paper is organized as follows: The second part provides a brief overview of relevant research works. In the third part, we present a detailed exposition of our proposed saliency prediction model. The fourth part focuses on our proposed saliency-based HRI framework. Moving forward, the fifth part delves into presenting the experimental setup and results acquired. In the final section, we conclude the paper by summarizing our findings and discussing potential avenues for future research.

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 [32]. 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 [21, 33]. 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" [22, 34–39]. 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. [40], Lee et al. [33] and Li and Yu [21] Wang et al. [41] and Zhang et al. [42] [23–25, 43–45]. 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 [23].

Static image saliency research is almost mature. However, subsequent trials to employ these models on video show a reduced performance [46]. 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 [13, 46–53]. Some of these studies [46, 47, 49] 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 [54], background subtraction [55], optical flow [56] 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 [29, 57] that accounts for color images and motion fields separately, or two-layer LSTM with object information [58, 59]

As one of the first attempts, [29] 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.

[58] 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.

[57] 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.

[59] 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 [60], 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 [61], 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 [62], 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.

[63] 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 [64], 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, [65] applied LSTMs to predict video saliency maps, relying on both short- and long-term memory of attention deployment.

In [66], 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 [67], 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.

[30], 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.

[68] 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 [40, 69–74] 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 [75–78]. These dataset improved the understanding of human visual attention and boosted the performance of computational models.

The DHF1K [30] 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 [77] 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 [79], 84.5 fixations Hollywood-2 dataset are located around the faces.

The UCF Sports dataset [77] 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 [80]. According to [79], 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 [23, 75, 76, 78, 81], or collected for a special purpose (e.g., salient objects in videos [72]). More importantly, none of the aforementioned datasets includes a preserved test set for avoiding potential data overfitting, which has seriously hampered the research process.

2.3 Saliency Based HRI Frameworks

Saliency-based human-robot interaction frameworks have gained significant attention in recent years, driving advancements in robot perception, attention allocation, and interaction capabilities.

For instance, [3] proposed an attentional mechanism for HRI, integrating saliency prediction models into a humanoid robot. The framework employed a combination of bottom-up saliency maps and top-down attention mechanisms to guide the robot's perception and attention allocation. The study demonstrated improved interaction capabilities, allowing the robot to respond to salient human cues effectively. In [1], an evaluation framework for visual attention in HRI is introduced. They incorporated saliency prediction models to estimate the saliency of objects in the robot's visual field. By leveraging saliency cues, the robot exhibited more natural and socially acceptable attention behaviors. The study highlighted the importance of incorporating human-like attention mechanisms for effective HRI.

An interactive HRI framework that utilized saliency-based perception to enhance robot understanding of human intentions in developed in [4]. By integrating saliency prediction models with intention recognition algorithms, the robot could identify and respond to human actions in real-time. The framework demonstrated improved interaction performance and enhanced robot comprehension of human behavior.

A research in [6] proposed a saliency-based HRI framework specifically for collaborative tasks. The framework integrated saliency prediction models with collaborative task planning algorithms, enabling the robot to allocate attention to relevant areas during cooperative activities. The study showcased the effectiveness of incorporating saliency-based perception in facilitating seamless collaboration between humans and robots.

A bio-inspired saliency-based attention model for HRI is developed in [2]. The framework utilized computational models inspired by human visual attention to guide the robot's attention allocation. By incorporating saliency prediction, the robot could attend to salient regions in the environment, resulting in more efficient and engaging interactions with humans.

A gaze-based HRI framework that utilized saliency prediction to infer the robot's focus of attention is developed in [82]. The framework incorporated eye-tracking technology and saliency maps to estimate the saliency of objects in the environment. The study showcased the benefits of using saliency-based gaze estimation for more natural and intuitive HRI.

In conclusion, the reviewed works in video saliency prediction and saliency-based human-robot interaction (HRI) frameworks highlight the importance of incorporating saliency prediction models for improved robot perception and interaction. Existing frameworks have demonstrated the effectiveness of saliency-based approaches but face limitations in adapting to dynamic environments and the lack of representative benchmarks. To address these challenges, we propose a novel approach that leverages advanced techniques, such as stacked-ConvLSTM architecture and XY-Shift frame differencing, to enhance saliency prediction in HRI settings.

3 Video Saliency Prediction Model

We present a novel video saliency prediction model based on stacked-ConvLSTM architecture. The architecture of our model is illustrated in Figure 1. It combines the power of convolutional and recurrent networks to effectively capture spatio-temporal information. To preprocess the input data, we introduce a novel XY-shift frame differencing layer. This layer computes the absolute difference between an image and its shifted copy, producing a high-pass filtered map. Additionally, we employ a three-frame differencing method to incorporate temporal information into the spatial domain. By magnifying the impact of temporal features, this technique enhances the capability of the stacked-ConvLSTM component in spatio-temporal saliency prediction. Consequently, our model achieves accurate video saliency prediction with improved generalization.

In this section, we provide a comprehensive description of our proposed model architecture and its three crucial components: the stacked-ConvLSTM module, the VGG-16 network [31], and the XY-shift frame differencing module.

3.1 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 [83], XY-shift frame differencing and the stacked-ConvLSTM component.

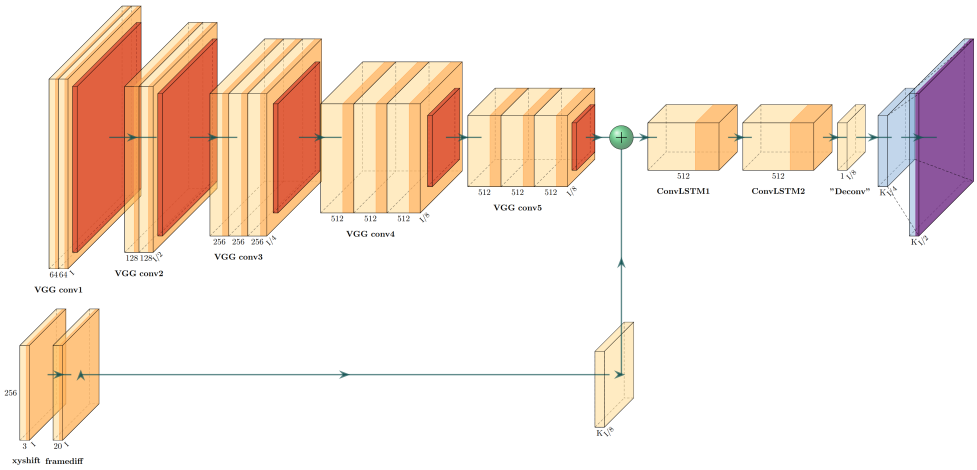


Fig. 1: Interactive Video Saliency Identification With Attentive ConvLSTM Architecture

3.2 Implementation Details

The implementation details of our model are as follows. Initially, two streams of data are fed into both the VGG-16 network and the frame differencing components. The VGG-16 network is responsible for extracting spatial features from the raw image frames [31]. To preserve more

spatial details, we remove the Pool 4 and Pool 5 layers, resulting in a downsampling factor of x8 instead of x32. At each time step t, the size of the input RGB image Xt is (224x224x3), and the output characteristic size of this component is [32, 40, 512].

Simultaneously, we apply batch-level XY-shift frame differencing and three-frame differencing to magnify temporal features in the spatial domain for each member of the batch. The XY-shift frame differencing operation calculates the difference between a frame and its shifted replica. This operation effectively acts as a high-pass filter, but with significantly lower computational resource requirements. Our primary motivation for using this method is to reduce the visibility of irrelevant background objects and highlight 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.

Following the XY-shift frame differencing, we employ an enhanced three-frame differencing technique. This technique utilizes the output of the XY-shift differencing operation. It involves taking three consecutive frames and computing the pixel-wise differences between the current frame and the previous frame, as well as between the current frame and the next frame. Subsequently, we extract the maximum pixel-wise values from these two resulting frames. This improved three-frame differencing method is specifically designed to enhance the extraction of temporal features from datasets in the spatio-temporal domain.

The formalization of the improved three-frame differencing method is described as follows in Equation 2. Let A, B, and C represent the first channel of three consecutive XY-shift frame differenced frames with a shape of (h, w). Furthermore, let B denote 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 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:

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

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

Following the aforementioned components, we incorporate a residual layer that integrates the spatial features extracted by VGG-16 and the output frames from the frame differencing

step. This residual layer aims to fuse the information from both sources into a unified representation. The resulting output of the residual layer is a [30x40x512] feature map, which is subsequently fed into our stacked-ConvLSTM network.

The stacking of ConvLSTM modules is crucial to accommodate greater model complexity. Despite the existence of large-scale datasets such as DHF1K, which contain 1K videos, the available training data is still limited, considering the high inter-frame correlation within the same video [58, 84]. By increasing the complexity of the model, we can capture more intricate features, thus yielding a robust video saliency prediction model. The feature map size after the stacked-ConvLSTM module is 32x40x256.

To obtain saliency maps corresponding to different loss functions employed in this research, we further process the output of the stacked-ConvLSTM module. This involves passing the feature map through a convolutional layer with a kernel size of 1x1, followed by upsampling. The resulting saliency maps have dimensions of 128x160x1 and 64x80x1, respectively.

3.3 Loss Functions

To better generate robust saliency maps, we use three loss functions as used in [85] and [30]. Linear Correlation Coefficient(CC) [86], the Kullback-Leibler divergence (KLD) [87] and Normalized Scanpath Saliency (NSS) [88]. 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:

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

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. αs 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 $\hat{S}$:

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

Normalized Scanpath Saliency metric was introduced in [88], 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:

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

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

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

3.4 Training Protocol

During the training process, our model is trained iteratively using sequential fixation and image data. The training procedure involves cascading a video training batch with an image training batch.

In the video training batch, a loss function is defined over the final dynamic saliency prediction obtained from the LSTM module. For each video training batch, we select 20 consecutive frames from the same video. Both the video and the starting frame within the video are randomly chosen, ensuring diversity in the training data.

In the image training batch, we set the batch size to 20, and random sampling is employed to select images from an existing static fixation dataset. This allows us to incorporate static saliency information into the training process and further enhance the model's performance.

By combining both sequential fixation and image data in the training procedure, our model learns to effectively predict saliency in dynamic video sequences while also leveraging the knowledge gained from static fixation datasets.

4 HRI Framework

4.1 Overview

Our saliency-based Human-Robot Interaction framework represents a cutting-edge system that utilizes the findings and advancements derived from our saliency prediction research to significantly enhance the interaction between humans and robots. The primary objective of our framework is to establish a more natural and intuitive communication channel between these two entities. By incorporating the valuable insights gained from our video saliency prediction model, we can effectively identify and prioritize visually salient regions within a stream of images or video frames, enabling the robot to better understand and respond to human visual attention.

In Figure 2, we provide a visual representation of our comprehensive HRI framework. It encompasses various crucial components that synergistically work together to seamlessly integrate the saliency prediction capabilities into the HRI environment. These components include the perception module, which incorporates the video saliency prediction model, the human-robot interaction manager, and the behaviour manager modules.

At the core of our system lies a powerful video saliency prediction model designed specifically to detect and emphasize crucial regions within the video feed, thereby enabling the robot to comprehend the user's focal points more effectively. This model is a direct outcome of our extensive research in video saliency prediction.

The human-robot interaction manager assumes a pivotal role in orchestrating the interaction dynamics between the robot and the human user. Built upon the robust Robot

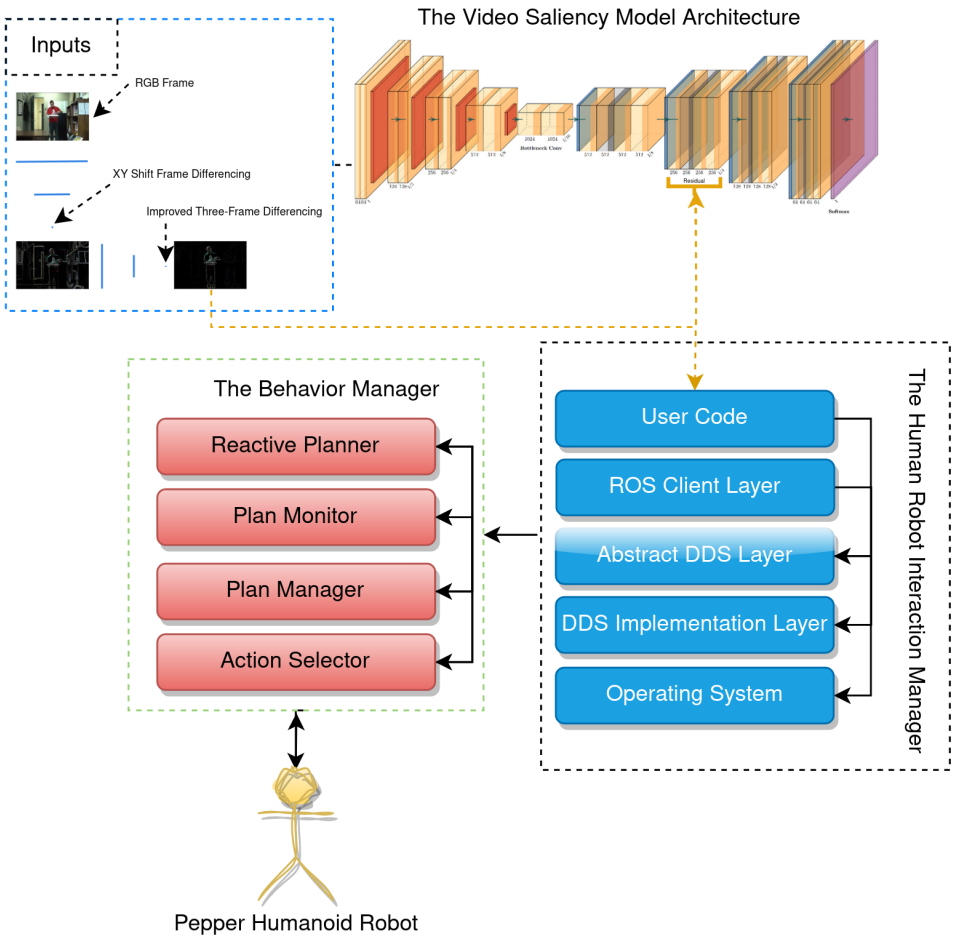


Fig. 2: Saliency-based human-robot interaction framework

Operating System (ROS) framework [89], it facilitates seamless communication and coordination among the different modules, robots, and humans involved in the interaction process. Leveraging the capabilities of ROS, our framework enables the robot to receive commands, instructions, and cues from the user and respond to them in a natural and coherent manner.

Working in tandem with the human-robot interaction manager, the behavior manager takes charge of regulating the robot's behavior based on the insights provided by the saliency prediction model. Leveraging the rich information extracted from the video saliency analysis, the behavior manager makes informed decisions about the robot's responses to the user's actions and ensures that these responses align with the user's expectations and preferences. By dynamically adjusting the robot's behavior based on the saliency predictions, our system strives to create a harmonious and intuitive interaction experience.

Through the integration of the video saliency prediction model, the human-robot interaction manager, and the behavior manager, our system establishes a coherent framework that enhances the human-robot interaction. The video saliency prediction model serves as a critical component, empowering the robot with a deeper understanding of the user's areas of interest. The human-robot interaction manager, facilitated by ROS, enables seamless communication and coordination, while the behavior manager ensures that the robot's actions are aligned with the user's attentional cues and intentions.

4.2 Implementation Details

The developed HRI framework was implemented and evaluated in two distinct environments: a virtual environment utilizing the Gazebo simulation software and a physical environment employing the Pepper robot. These environments served as testbeds to assess the effectiveness and performance of the framework in different settings. Here, we will delve into the details of each implementation.

4.2.1 Simulation Environment

To conduct the experiment, we employed a 3D model of the Pepper robot, a popular humanoid robot, in the Gazebo environment. Gazebo provided a highly realistic simulation platform that allowed us to evaluate the effectiveness of the video saliency prediction model under controlled conditions. The experimental workflow encompassed several key steps. Firstly, we set up the Gazebo environment, configuring the virtual world to closely resemble real-world scenarios. Next, we implemented the computer vision models as ROS nodes, leveraging the ROS framework for seamless communication between different modules, robots, and humans. This integration enabled the video saliency prediction model to process the visual input from the virtual robot's camera. Furthermore, we developed a behavior manager module within the framework to govern the camera movements of the virtual robot. This module utilized the predictions from the video saliency model to determine the most relevant areas of interest and control the robot's focus accordingly. Finally, we meticulously evaluated the system's performance across a range of diverse scenarios in the virtual environment. By conducting experiments in this simulated setting, we were able to iteratively test and refine our human attention models, optimizing their accuracy and robustness before deploying them on a physical Pepper robot in real-world HRI scenarios.

4.2.2 Real Environment

The integration of the Pepper robot with the ROS was achieved through the utilization of the ROSBridge protocol. This protocol facilitated smooth communication between the robot and external systems, enabling seamless integration of the Pepper robot with ROS. The integration process was made possible by leveraging the Naoqi_driver, which provided the necessary interface for the Pepper robot to interact with ROS and utilize its extensive range of tools and functionalities. The integration itself relied on the ROS message protocol, ensuring standardized data exchange in a format native to ROS. This approach offers several advantages, including a streamlined and efficient communication process, as well as enhanced compatibility with a wide array of ROS-based tools and functionalities.

In the development of our saliency based HRI framework, we leveraged ROS nodes to handle various aspects, such as computer vision, navigation, and the attention model itself. The human attention model utilized input from the robot's vision sensors to determine the appropriate focus of the robot's attention. This information was then communicated to the robot's actuator, specifically the wheel motors, to facilitate appropriate movement of the robot's body. To establish communication with the Pepper robot, we designed a ROS package compatible with both ROS Noetic and ROS2 Humble versions, utilizing the Naoqi driver. This package acquired image information in RGB format through the `/image_raw` topic and sent linear and angular velocity commands to the `/cmd_vel` controller. Through this approach, we achieved seamless integration of our saliency prediction model with the Pepper robot, enabling more intuitive and adaptive interactions between the robot and its environment.

5 Results

5.1 Saliency Prediction Model

5.1.1 Overview

The saliency prediction model results section focuses on evaluating the performance of our saliency prediction model. We present a comprehensive analysis of the model's effectiveness in predicting saliency in videos. To assess the model's performance, we conducted experiments using a large-scale video saliency dataset, DHF1K. We compared our model against state-of-the-art dynamic and static video saliency prediction models, employing various evaluation metrics. Additionally, we examined the impact of different loss functions on the model's performance. In the following sections, we will delve into a detailed analysis of the results obtained from our saliency prediction model. We will discuss the performance of the model across different evaluation metrics and provide visual comparisons with other state-of-the-art models.

5.1.2 Datasets

We use the DHF1K [30] dataset for training and evaluation. We use only the first 70% of the DHF1K dataset and used 60%/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.

5.1.3 Evaluation Metrics

We use the following evaluation metrics

Normalized Scanpath Saliency

$$\mathrm{NSS} = \frac{(S-\mu)}{\sigma}$$

Where:

S represents the saliency value at the fixation location.

μ denotes the mean saliency value across all possible fixation locations.

σ represents the standard deviation of the saliency values across all possible fixation locations.

Similarity Metric

$$\text{Similarity}(X,Y) = \frac{\sum_{i=1}^{n} w_i \cdot X_i \cdot Y_i}{\sqrt{\sum_{i=1}^{n} w_i \cdot X_i^2} \cdot \sqrt{\sum_{i=1}^{n} w_i \cdot Y_i^2}}$$

where:

$\text{Similarity}(X,Y)$ represents the similarity measure between variables X and Y.

X_i and Y_i denote the individual observations of variables X and Y, respectively.

w_i represents the weight associated with each observation, which can be assigned based on domain knowledge or other considerations.

$\sum_{i=1}^{n}$ denotes the summation over the range of i from 1 to n, where n is the number of observations.

Linear Correlation Coefficient

$$\rho_{X,Y} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

where:

$\rho_{X,Y}$ represents the correlation coefficient between variables X and Y.

X_i and Y_i denote the individual observations of variables X and Y, respectively.

$\bar{X}$ and $\bar{Y}$ represent the means of variables X and Y, respectively.

$\sum_{i=1}^{n}$ denotes the summation over the range of i from 1 to n, where n is the number of observations.

AUC-Judd

$$\text{AUC Judd} = \frac{\sum_{\tau}(F(\tau) - F(\tau^-)) \cdot s(\tau)}{\sum_{\tau}(1 - F(\tau^-))}$$

where:

AUC Judd represents the Area Under the Curve (AUC) for the Judd metric.

τ denotes the threshold values at which the fixation prediction map is binarized.

$F(\tau)$ represents the cumulative distribution function (CDF) of the saliency map values.

$F(\tau^-)$ represents the CDF values at τ^-, which is the threshold value just below τ.

$s(\tau)$ represents the linearized saccadic eye movement density at threshold τ.

Shuffled AUC

$$\text{Shuffled AUC} = \frac{1}{N \cdot M} \sum_{i=1}^{N} \sum_{j=1}^{M} \left(R(x_i, y_j) - \frac{1}{2} \right)$$

where:

Shuffled AUC represents the evaluation metric for comparing two sets of ranked data.

N represents the number of samples in the first set.

M represents the number of samples in the second set.

$R(x_i, y_j)$ represents the ranking of the ith sample in the first set relative to the jth sample in the second set.

5.1.4 Competitors

To prove the effectiveness of our proposed model, we compare our model with sixteen saliency models. Among them, [30], PQFT [13], Seo et al. [49], Rudoy et al.[48], Hou et al. [50], Fang et al. [51], OBDL [52], AWS-D [53], OM-CNN [58], and Two-stream [29] are dynamic saliency models. Furthermore, ITTI [21], GBVS [22], SALICON [23], DVA [24], Shallow-Net [25], and Deep-Net [25] 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.

5.1.5 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.

5.1.6 Performance Comparison

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 [30]. This is directly attributed to the novel XY-shift frame differencing technique and stacked-ConvLSTM network incorporated in our architecture.

5.1.7 Analysis

In the course of our research, we have conducted extensive experiments. Here, we analyse our model and 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 [24], Deep-Net [25] and dynamic OM-CNN [58], Two-stream [29], ACL [30] 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 [21],GBVS [22]. A significant performance degradation is observed when static saliency prediction algorithms are employed for dynamic saliency prediction problem sets PQFT [13], [49], [48], [50], [51]. 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.

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

	Models/Datasets	DHF1K				
		AUC-J	**SIM**	**s-AUC**	**CC**	**NSS**
Dynamic models	[13]	0.699	0.139	0.562	0.137	0.749
	[49]	0.635	0.142	0.499	0.070	0.334
	[48]	0.769	0.214	0.501	0.285	1.498
	[50]	0.726	0.167	0.545	0.150	0.847
	[51]	0.819	0.198	0.537	0.273	1.539
	[52]	0.638	0.171	0.500	0.117	0.495
	[53]	0.703	0.157	0.513	0.174	0.940
	[58]	0.856	0.256	0.583	0.344	1.911
	[29]	0.834	0.197	0.581	0.325	1.632
	[30]	**0.885**	**0.311**	0.553	**0.415**	**2.259**
Static models	[21]	0.774	0.162	0.553	0.233	1.207
	[22]	0.828	0.186	0.554	0.283	1.474
	[23]	0.857	0.232	0.590	0.327	1.901
	[25] Shallow-Net	0.833	0.182	0.529	0.295	1.509
	[25] Deep-Net	0.855	0.201	0.592	0.331	1.775
	[24]	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**

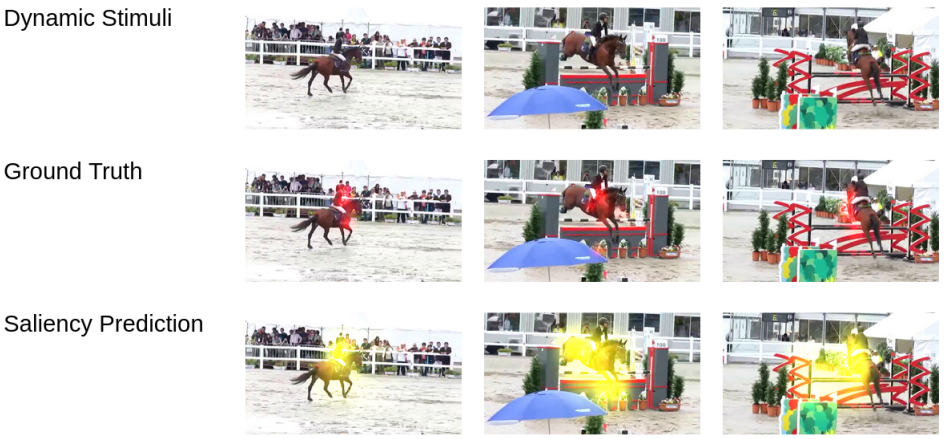


Fig. 3: Qualitative results of our video saliency model on DHF1K Dataset

5.1.8 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 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 [90] 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.

5.2 Saliency based HRI Framework

5.2.1 Overview

The saliency based HRI framework results section presents a comprehensive analysis and evaluation of the saliency-based HRI framework. This framework combines the advancements in saliency prediction models with human-robot interaction to create a more natural and intuitive interaction between robots and humans.

5.2.2 Simulation Environment

We evaluate our framework on the Gazebo platform using subjective rating evaluation criteria to assess the intuitiveness of human-robot interaction based on our human attention models. This evaluation involves seven students from Addis Ababa University.

During the evaluation of our HRI framework, we involved professionals from various backgrounds, including two Software Engineers, three Artificial Intelligence Experts, and two Cyber Security professionals. The diverse backgrounds of our evaluators allowed for a comprehensive evaluation from different perspectives. Participants evaluated our framework based on criteria such as intuitiveness, engagement, trust, and user satisfaction across four controlled experimentation scenarios. These scenarios included (1) a moving object scenario, (2) an intuitive human interaction scenario, (3) an anthropomorphic scenario, and (4)

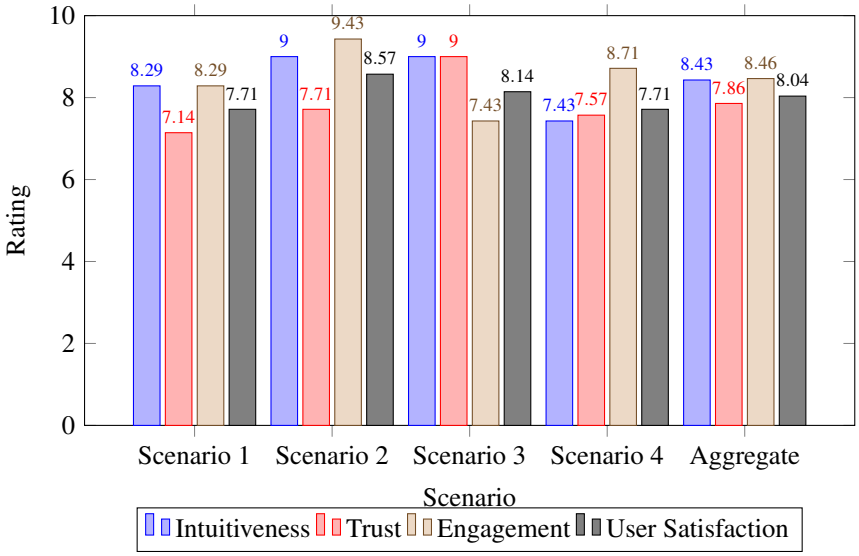


Fig. 4: Evaluation Ratings for Scenarios

a dynamic environment scenario. To provide a visual representation of these scenarios, we included screenshots in Figure 5a[1], Figure 5b[2], Figure 5c[3], and Figure 5d[4].

Consequently, we provide an aggregation of users subjective rating in Figure 4. According to the results, our human attention based anthropomorphic human-robot interaction framework enabled intuitive, trustworthy, and fairly satisfying interaction capabilities with the virtual environment.

5.2.3 Real-Time Embedded Strategy

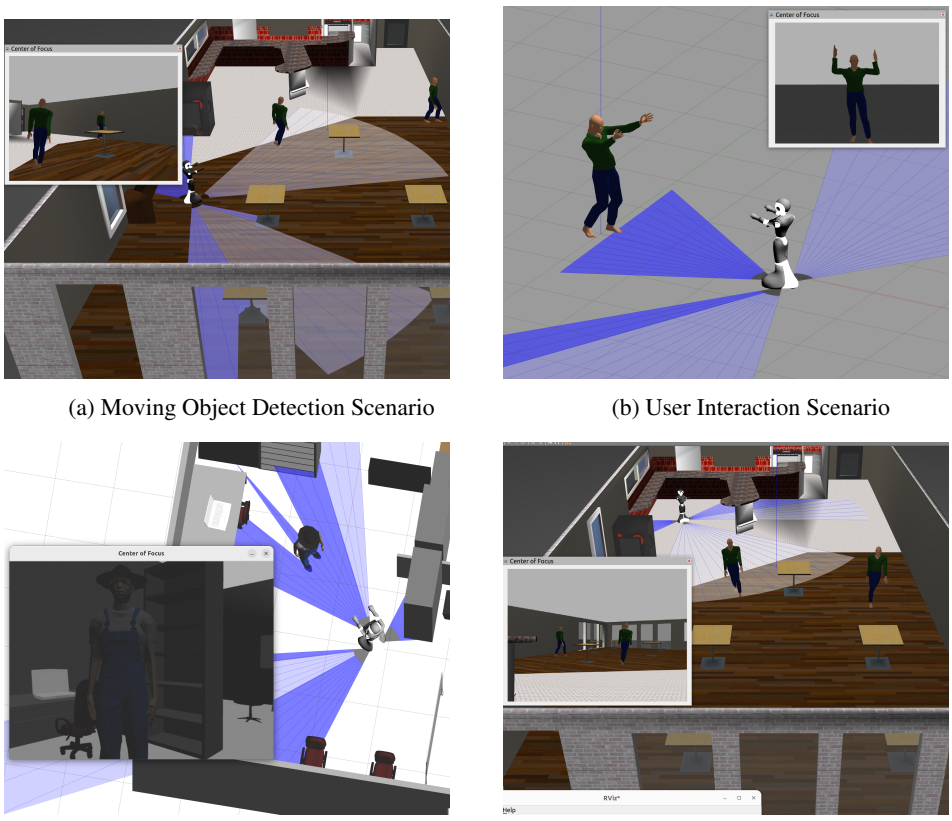
Although we conducted an extensive evaluation in a simulated environment, we also performed a qualitative analysis of our HRI framework using a Pepper humanoid robot. The tests we conducted involved moving object detection, user interaction, and anthropomorphic robot action using both saliency prediction model.

To demonstrate the efficiency of our approach in the real-time and low-resources constraints of the RoboCup@Home, we employed ourselves to optimize and deploy the architecture onboard the Pepper robot. To do so, models weights are first converted to Tensorflow Lite [5] format for lightweight inference on the robot CPU [6]. Then, a synchronizer is added to ensure low-latency communication between the attention module, moving object module, and behavior module.

Figure 6 displays a screenshot of the Pepper humanoid robot operating with our saliency based HRI framework.

---

[1] Pepper paying attention to moving bodies
[2] Pepper interacting with temporarily salient body
[3] Pepper acting humanly in a still environment
[4] Pepper attending a very dynamic environment and on the move
[5] https://www.tensorflow.org/lite/guide?hl=en
[6] Intel Atom™ E3845 @ 1.91GHz x 4



(a) Moving Object Detection Scenario

(b) User Interaction Scenario

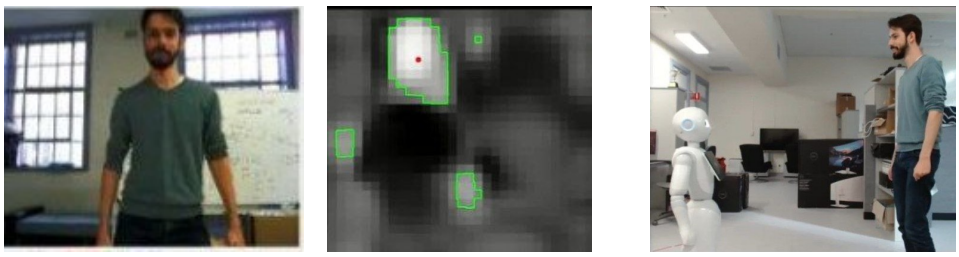
(c) Anthropomorphic Robot Behaviour Scenario

(d) Dynamic space Interaction Scenario

Fig. 5: Pepper Humanoid Robot Operating in Simulated Environment

6 Conclusion

In this research, we present a comprehensive study that merges two key areas: dynamic saliency prediction and human-robot interaction. We propose a novel deep learning-based dynamic saliency prediction model that leverages a unique XY-shift frame differencing technique and a stacked-ConvLSTM network. Extensive experimentation on the DHF1K dataset showcases the effectiveness and superiority of our model compared to 15 state-of-the-art models, while also demonstrating its competitiveness against the leading dynamic saliency prediction model [30]. Besides, we introduce a saliency based HRI framework. The framework is evaluated through subjective ratings from participants across various scenarios, assessing intuitiveness, trust, engagement, and user satisfaction. Moreover, we evaluted the average latency of our framework taking the time it takes for the robot to commit instructions. Extensive testing is conducted in both simulated and physical environments, utilizing the Pepper humanoid robot. By incorporating cutting-edge saliency prediction technique and HRI design principles, our framework enables robots to seamlessly adapt their behavior based on human attention cues.



(a) Pepper view (b) Real-time saliency (c) Global view

Fig. 6: Pepper Humanoid Robot Operating in the Real World: Human Interaction Scenario.

This interdisciplinary work represents a significant step forward in developing more efficient and intuitive HRI systems. It not only advances the field of dynamic saliency prediction but also contributes to the seamless integration of saliency prediction models into robot behavior. We also highlight potential research avenues for future exploration in HRI and human attention models. Overall, this research fosters the development of advanced HRI systems that better understand and respond to human attention, leading to more natural and effective interactions between humans and robots.

7 Conflict of Interest

On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

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

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

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

[4] Diaz, C.G., Perry, P., Fiebrink, R.: Interactive machine learning for more expressive game interactions. In: 2019 IEEE Conference on Games (CoG), pp. 1–2 (2019). IEEE

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

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

(2018)

[7] Chen, Y., Zhang, W., Wang, S., Li, L., Huang, Q.: Saliency-based spatiotemporal attention for video captioning. In: 2018 IEEE Fourth International Conference on Multimedia Big Data (BigMM), pp. 1–8 (2018). IEEE

[8] Wang, H., Xu, Y., Han, Y.: Spotting and aggregating salient regions for video captioning. In: Proceedings of the 26th ACM International Conference on Multimedia, pp. 1519–1526 (2018)

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

[10] Lateef, F., Kas, M., Ruichek, Y.: Saliency heat-map as visual attention for autonomous driving using generative adversarial network (gan). IEEE Transactions on Intelligent Transportation Systems (2021)

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

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

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

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

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

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

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

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

[19] Xie, Y., Lu, H.: Visual saliency detection based on bayesian model. In: 2011 18th IEEE International Conference on Image Processing, pp. 645–648 (2011). IEEE

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

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

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

[23] Huang, X., Shen, C., Boix, X., Zhao, Q.: Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 262–270 (2015)

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

[25] Pan, J., Sayrol, E., Giro-i-Nieto, X., McGuinness, K., O'Connor, N.E.: Shallow and deep convolutional networks for saliency prediction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 598–606 (2016)

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

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

[28] Rice, L., Wong, E., Kolter, Z.: Overfitting in adversarially robust deep learning. In: International Conference on Machine Learning, pp. 8093–8104 (2020). PMLR

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

[30] Wang, W., Shen, J., Guo, F., Cheng, M.-M., Borji, A.: Revisiting video saliency: A large-scale benchmark and a new model. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4894–4903 (2018)

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

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

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

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

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

[36] Judd, T., Ehinger, K., Durand, F., Torralba, A.: Learning to predict where humans look. In: 2009 IEEE 12th International Conference on Computer Vision, pp. 2106–2113 (2009). IEEE

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

[38] Yang, C., Zhang, L., Lu, H., Ruan, X., Yang, M.-H.: Saliency detection via graph-based manifold ranking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3166–3173 (2013)

[39] Jiang, H., Wang, J., Yuan, Z., Wu, Y., Zheng, N., Li, S.: Salient object detection: A discriminative regional feature integration approach. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2083–2090 (2013)

[40] Hou, Q., Cheng, M.-M., Hu, X., Borji, A., Tu, Z., Torr, P.H.: Deeply supervised salient object detection with short connections. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3203–3212 (2017)

[41] Wang, T., Borji, A., Zhang, L., Zhang, P., Lu, H.: A stagewise refinement model for detecting salient objects in images. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4019–4028 (2017)

[42] Zhang, P., Wang, D., Lu, H., Wang, H., Ruan, X.: Amulet: Aggregating multi-level convolutional features for salient object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 202–211 (2017)

[43] Vig, E., Dorr, M., Cox, D.: Large-scale optimization of hierarchical features for saliency prediction in natural images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2798–2805 (2014)

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

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

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

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

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

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

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

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

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

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

[54] Mech, R., Wollborn, M.: A noise robust method for segmentation of moving objects in video sequences. In: 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 4, pp. 2657–2660 (1997). IEEE

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

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

[57] Zhao, T., Wu, X.: Pyramid feature attention network for saliency detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3085–3094 (2019)

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

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

[60] Song, H., Wang, W., Zhao, S., Shen, J., Lam, K.-M.: Pyramid dilated deeper convlstm for video salient object detection. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 715–731 (2018)

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

[62] Li, G., Xie, Y., Wei, T., Wang, K., Lin, L.: Flow guided recurrent neural encoder for video salient object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3243–3252 (2018)

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

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

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

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

[67] Gorji, S., Clark, J.J.: 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, pp. 7501–7511 (2018)

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

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

[70] Achanta, R., Hemami, S., Estrada, F., Susstrunk, S.: Frequency-tuned salient region detection. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1597–1604 (2009). IEEE

[71] Cheng, M.-M., Mitra, N.J., Huang, X., Torr, P.H., Hu, S.-M.: Global contrast based

salient region detection. IEEE transactions on pattern analysis and machine intelligence **37**(3), 569–582 (2014)

[72] Wang, W., Shen, J., Porikli, F.: Saliency-aware geodesic video object segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3395–3402 (2015)

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

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

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

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

[77] Mathe, S., Sminchisescu, C.: 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 (2014)

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

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

[80] Rodriguez, M.D., Ahmed, J., Shah, M.: Action mach a spatio-temporal maximum average correlation height filter for action recognition. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2008). IEEE

[81] Bylinskii, Z., Judd, T., Borji, A., Itti, L., Durand, F., Oliva, A., Torralba, A.: Mit saliency benchmark. MIT Press (2015)

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

[83] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255 (2009). Ieee

[84] Liu, N., Han, J., Zhang, D., Wen, S., Liu, T.: Predicting eye fixations using convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 362–370 (2015)

[85] Jiang, L., Xu, M., Liu, T., Qiao, M., Wang, Z.: Deepvs: A deep learning based video saliency prediction approach. In: Proceedings of the European Conference on Computer Vision (eccv), pp. 602–617 (2018)

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

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

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

[89] Quigley, M., Conley, K., Gerkey, B., Faust, J., Foote, T., Leibs, J., Wheeler, R., Ng, A.Y., *et al.*: Ros: an open-source robot operating system. In: ICRA Workshop on Open Source Software, vol. 3, p. 5 (2009). Kobe, Japan

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