Computers & Graphics 106 (2022) 200–209

Contents lists available at ScienceDirect

Computers & Graphics

journal homepage: www.elsevier.com/locate/cag

Special Section on CEIG 2022

SST-Sal: A spherical spatio-temporal approach for saliency prediction in 360° videos

Edurne Bernal-Berdun *, Daniel Martin, Diego Gutierrez, Belen Masia

Universidad de Zaragoza - I3A, Spain

ARTICLE INFO

Article history:
Received 12 May 2022
Accepted 3 June 2022
Available online 11 June 2022

Keywords:
Virtual reality
Scene Understanding
Saliency

ABSTRACT

Virtual reality (VR) has the potential to change the way people consume content, and has been predicted to become the next big computing paradigm. However, much remains unknown about the grammar and visual language of this new medium, and understanding and predicting how humans behave in virtual environments remains an open problem. In this work, we propose a novel saliency prediction model which exploits the joint potential of spherical convolutions and recurrent neural networks to extract and model the inherent spatio-temporal features from 360° videos. We employ Convolutional Long Short-Term Memory cells (ConvLSTMs) to account for temporal information at the time of feature extraction rather than to post-process spatial features as in previous works. To facilitate spatio-temporal learning, we provide the network with an estimation of the optical flow between 360° frames, since motion is known to be a highly salient feature in dynamic content. Our model is trained with a novel spherical Kullback–Leibler Divergence (KLDiv) loss function specifically tailored for saliency prediction in 360° content. Our approach outperforms previous state-of-the-art works, being able to mimic human visual attention when exploring dynamic 360° videos.

1. Introduction

Virtual reality (VR), which has been usually associated with the entertainment industry, has proven to be a powerful tool in many other fields such as the manufacturing industry, online marketing, architecture, design, or education. Although many software and hardware advances have been made in VR, this new technology presents challenges and limitations that are still open problems. Unlike traditional media, where the whole content is usually shown on a flat screen, VR content occupies the 360° space around the user. As in real life, only the part of the scene which falls into the observer's field of view (FoV) is seen, and it is by eye and head movements that the remainder of the environment can be explored. Therefore, *the user takes control of the camera*, choosing what to observe. Due to this paradigm shift, the patterns and visual behaviors known for 2D images or videos do not necessarily hold for 360° immersive environments.

As a fundamental step towards understanding how humans behave in VR environments, many approaches have focused on modeling which points in a scene are most likely to attract users' attention. They resorted to *saliency* as a topological measure of the conspicuity of the elements of that scene, or in other words, the probability of each element to receive a fixation from the observer. The study of fixation data from real observers gathered with eye-tracking devices has shown that there is an inter- and intra- observer variability when exploring 360° visual stimuli [1–3]. This variability makes the task of visual human attention prediction challenging, but, even though the behavior of multiple observers in response to the same stimulus is rarely the same, they all share some common, inherent patterns [1]. Understanding and modeling these patterns, which are usually influenced by the saliency of the scene, can enhance many applications in computer graphics, including foveated rendering [4–6] or image compression [2,7], or content generation [1,8,9], among others.

In this work, we present SST-Sal, a novel saliency prediction model for 360° videos. Our model is based on Recurrent Neural Networks (RNNs), specifically on Long Short-Term Memory (LSTMs) cells. LSTMs allow inferring temporal relationships between frames, a crucial aspect in dynamic content, where the observer's attention is influenced by what has previously happened in the 360° video. However, the traditional operations

* Corresponding author.
E-mail addresses: edurnebernal@unizar.es (E. Bernal-Berdun), danims@unizar.es (D. Martin), diegog@unizar.es (D. Gutierrez), bmasia@unizar.es (B. Masia).
URLs: https://edurnebernal.github.io/ (E. Bernal-Berdun), http://webdiis.unizar.es/~danims/ (D. Martin), http://giga.cps.unizar.es/~diegog/ (D. Gutierrez), http://webdiis.unizar.es/~bmasia/ (B. Masia).

https://doi.org/10.1016/j.cag.2022.06.002
0097-8493/

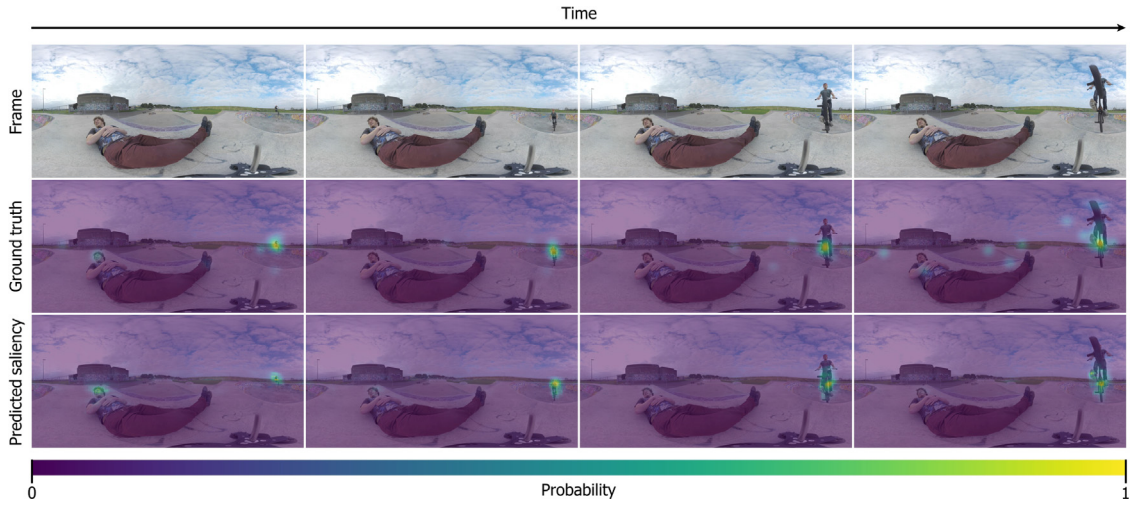

Fig. 1. In this work, we propose a saliency prediction network for 360° videos that follows an encoder–decoder architecture and is built over Spherical ConvLSTMs to leverage temporal information and account for the particularities of 360° content. We additionally provide our network with an optical flow estimation to facilitate spatio-temporal learning. Our model yields realistic results and outperforms previous state-of-the-art models by a large margin.

of LSTMs have been replaced in our model by spherical convolutions, presenting Spherical Convolutional LSTMs (ConvLSTMs). These spherical convolutions allow accounting for the distortion introduced when projecting the 360° frames onto a 2D plane, while extracting spatial features from the 360° videos. SST-Sal follows an encoder–decoder approach, in which both encoder and decoder are built over these Spherical ConvLSTMs cells, thus accounting for both temporal and spatial information at feature encoding and decoding time. This approach deviates from previous state-of-the-art works, where feature extraction is performed only with spatial information, neglecting the temporal relationships between frames.

Besides the 360° RGB frames represented with an equirectangular projection, we provide our model with optical flow estimations between consecutive frames, which has proved to enhance our saliency predictions. Optical flow can be considered a measure of the movement of the elements on a scene when the 360° videos are filmed with a static camera, and both the velocity and direction of the elements greatly affect saliency in dynamic content. Furthermore, we present a novel spherical Kullback–Leibler Divergence (KLDiv) loss function specifically tailored to 360° content.

Our contributions can be summarized as follows:

- We introduce an encoder–decoder architecture based on ConvLSTMs that, for the first time, accounts for temporal information when encoding and decoding the feature vectors.
- We show that using operations that are aware of the particularities of the equirectangular representation favors the correct interpretation of 360° content. Therefore, we present a saliency prediction network built over spherical convolutions and propose a novel spherical KLDiv loss function.
- We provide our model with optical flow estimations that allow our network to learn the relationships between motion and saliency, enhancing our saliency predictions.

All three contributions were validated through ablation studies and comparisons to state-of-the-art works. Our code and models are publicly available at: https://graphics.unizar.es/projects/SST-Sal_2022.

2. Related work

2.1. Modeling visual attention in traditional 2D content

In the last decades, many works have been devoted to saliency prediction in 2D content. The method proposed by Itti et al. [10] in 1998 can be considered the seminal work in this regard. They propose a heuristic approach in which they extract low-level features from the images, such as color, orientation, or high contrast areas, and linearly combine them to obtain a final saliency map. Several follow-up works [11–14] continued with this bottom-up strategy, exploring different hand-crafted features to improve the predicted saliency. However, these strategies based on low-level features often failed to capture the actual eye movements, thus novel approaches [15–17] arose to palliate this limitation, proposing to use machine learning techniques to combine different low-level, middle-level, and high-level image features.

With the emergence of deep neural networks (DNNs), numerous works [18–24] opted for data-driven approaches in the field of saliency prediction, leading to unprecedentedly accurate models that increasingly resembled human visual behavior. One of the first works to exemplify the superiority of DNNs was DeepFix [18], a fully convolutional neural network (CNN) that proved to surpass state-of-the-art results relying on previous approaches based on hand-crafted features. Later, work such as Martin et al.'s [25], not only show the advantages of using CNNs, but also exemplified the benefit of considering the temporal, and especially non-deterministic, nature of human visual attention employing architectures such as Bayesian ConvLSTMs.

Although saliency prediction in 2D images has been extensively explored in recent years, the body of literature specifically addressing this topic in dynamic content (i.e., video), is quite narrow. Works such as the ones from Bak et al. [26,27], or the one from Jiang et al. [28] also resorted to DNNs to extract spatial features, but these works were also designed to handle temporal information. Specifically, the two latter use recurrent neural networks (RNN), whereas the former obtains that information from optical flow estimations. However, these models are not suitable for saliency prediction in 360° videos as they do not account for the particularities of 360° content, such as the distortion introduced by the sphere-to-plane projection or the limited FoV.

2.2. Modeling and predicting attention in 360° images

The evolution of saliency prediction models for 360° content has been closely linked to that of traditional one, where heuristic approaches [29–34], established a baseline until the rise of DNNs. Most of the works proposed for spherical images that follow a data-driven approach [21,22,35,36] have been adapted from, or strongly based on 2D image saliency models. However, one of the main challenges when using 2D saliency models for 360° content is dealing with the distortion introduced by the sphere-to-plane projection. Most of the works tried to diminish the effect of distortion in their final predictions by using representations such as cube-maps or viewport images, that yield lower distortion than equirectangular projections, but still deform the images. Nonetheless, these representations based on multiple images lead to discontinuities, redundant image boundaries, and repeated computations, which can hamper the prediction process.

Some state-of-the-art works [37–39] propose a different strategy to alleviate this issue. Instead of manipulating the projection of the content, they modified the neural network structure to account for 360° content particularities. Haoran et al. [37] represent the 360° images with a Geodesic ICOsahedral Pixelation (GICOPix), presenting graphical convolutional networks (GCNs) as an alternative to traditional CNNs. Martin et al. [40] employed the spherical convolutions proposed in SphereNet [38], where the convolutional kernel is applied on the spherical space and then projected into the 2D plane. Consequently, the kernel used for convolution and pooling presents the same distortion as the equirectangular image. A similar approach is proposed by Yanyu et al. [39], but employing a circular crown kernel on the sphere instead of the traditional square kernel.

Although dynamic content consists of a succession of images and could be considered an extension of saliency prediction in 360° images, the human visual behavior towards it differs greatly from the static case. Aspects such as the movement of the elements or the plot of the sequence affect the viewer's attention, causing the saliency of each frame to be influenced by previous frames.

2.3. Modeling and predicting attention in 360° videos

Previous works such as PanoSalNet [41] followed a naïve approach, presenting saliency prediction models for 360° videos whose CNN-based estimation networks were similar to those used for images. However, predicting an independent saliency map for each frame does not take temporality into consideration. Due to this limitation, many posterior works resorted to a different approach, in which spatio-temporal features are extracted in order to make a more suitable prediction.

A frequent choice in state-of-the-art models is the use of RNNs, especially LSTMs architectures, given their ability to retain temporal information. Cheng et al. [42] use a cube-map representation with a novel cube padding technique applicable to CNNs, which solves the image boundary problem. With it, they obtain the prior saliency maps of each frame with a CNN-based network, and then pass them through a convolutional LSTM (ConvLSTM) to capture relations between saliency in consecutive frames, obtaining the definitive saliency prediction.

Dahou et al. proposed a more sophisticated architecture with ATSal [43], in which they combine the saliency maps obtained from two different streams (namely attention and expert). The attention stream is a convolutional encoder–decoder with an attention mechanism, which is intended to extract global static saliency from the equirectangular frames. On the other hand, the expert stream composed of two SalEMA [44] networks, based on CNNs and LSTMs, tries to learn spatio-temporal saliency information from a cube-map representation.

Xu et al. [39] proposed a network for 360° video saliency prediction that includes spherical convolutions, pooling, and Mean Squared Error (MSE) loss, but also implements a Spherical ConvLSTM. The model's architecture is based on a U-Net [45], but with the inclusion of a Spherical ConvLSTM in its bottleneck to capture the temporal information.

However, these previous works either ignore the temporal information [41], or neglect the temporal relationships between frames when extracting the spatial features. Therefore, our proposal is to extract the frames' spatial features with temporal awareness. Instead of just presenting a LSTM-based module in the bottleneck [39,43], or at the end of the model's pipeline [42] as in previous works, our encoder–decoder architecture is entirely built over Spherical ConvLSTMs. This allow us to obtain the truly relevant spatio-temporal features in the overall context of the 360° video. Moreover, as a novelty with respect to previous works, we provide our network with optical flow estimations. The movement of objects is a determining factor in human visual attention when observing dynamic content, thus these optical flow estimations allow our model to learn the possible relationships between motion and saliency.

3. A model for saliency prediction in 360° videos

In this Section, we first include an in-depth view of our model (Section 3.1) detailing its main features: Spherical ConvLSTMs and optical flow. Then, we present our novel spherical Kullback–Leibler Divergence loss function used to train our network (Section 3.2), and conclude with additional training details (Section 3.3).

3.1. Our model

Our saliency prediction model is built following an encoder–decoder architecture, which is tailored to the particularities of dynamic 360° content by leveraging spherical convolutional recurrent networks (Spherical ConvLSTMs), whose spatio-temporal nature endorses their use in our problem. Specifically, the encoder module, formed by a Spherical ConvLSTM and a spherical Max Pooling layer, extracts the spatio-temporal features from an input sequence of RGB-optical flow pairs. On the other hand, the decoder, built with a Spherical ConvLSTM and an Up-Sampling layer, leverages that extracted latent information to predict a sequence of saliency maps (see Fig. 2).

Optical Flow. The movement of objects, people, or animals in 360° videos influences human visual attention [46,47], and both their speed and direction can be relevant features to consider when predicting saliency. We thus resort to optical flow as a measure of the relative movement between the elements of a scene and the camera throughout a sequence, and input it to the network alongside the RGB frames.

In this work, we use the optical flow estimation provided by RAFT [48], a DNN trained on the Sintel Dataset [49], which has shown outstanding results on both synthetic and real images. Although RAFT is designed to work on traditional 2D images, the provided optical flow estimation was evaluated and found to be sufficiently accurate (see Section 4.4.4), since most of the motion is usually concentrated in the less distorted region of equirectangular images — the equator.

We feed our model with the concatenation of a sequence $\mathcal{S}$ of frames f with length n, and the corresponding sequence $\mathcal{O}$ of optical flow estimations o. Therefore, the model's input $\mathcal{I}$ can be defined as:

$$\mathcal{I} = [\mathcal{S}, \mathcal{O}] = \{f_0\ o_0, \ldots, f_t\ o_t, \ldots, f_n o_n\} \qquad t \in [0, n], \tag{1}$$

where f and o are RGB equirectangular images of size WxHx3 (W and H are the width and height of the frames), both with

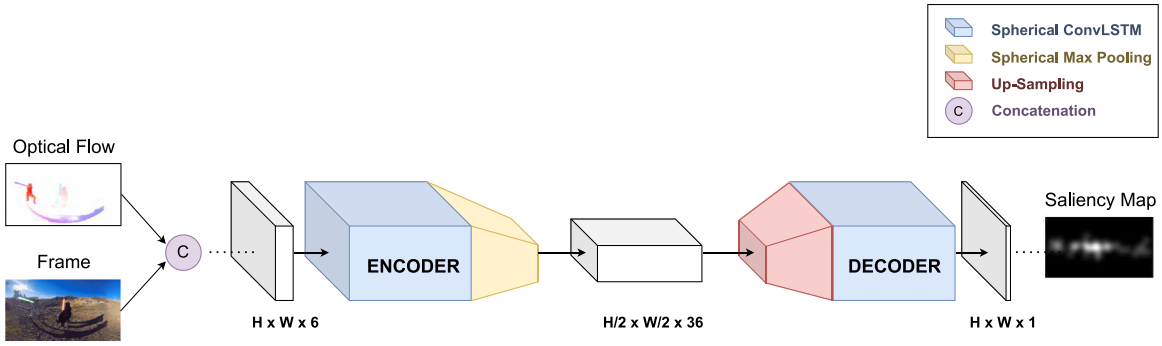

Fig. 2. Network's encoder–decoder architecture: The encoder is composed of a Spherical ConvLSTM and a spherical Max Pooling layer. It takes as input the concatenation of the RGB image (frame) and the optical flow. The decoder, formed by a Spherical ConvLSTM and an Up-Sampling layer, decodes the feature vector predicting the final saliency. Note that the recurrence of our model is not illustrated in the figure for the sake of clarity, thus only one timestep t is represented.

three channels. The timestep of each element on the sequence is represented by t. After performing the corresponding ablation study (see Section 4.4.1), we selected $W = 320$ and $H = 240$ as the input spatial resolution, and empirically set $n = 20$.

Spherical ConvLSTM. Dynamic videos contain elements that are presumably to move, change, or disappear. These continuous variations in position, appearance, or illumination are likely to attract the observers' attention, hence being perceived as salient. ConvLSTM cells can extract and process these temporal features from sequential data (e.g., videos) and infer temporal relations between them [50]. In ConvLSTMs, the fully-connected structures from traditional LSTMs are replaced by convolutional operations, which are able to handle spatial features over time. The equations defining our ConvLSTMs are as follows:

$$\begin{aligned} i_t &= \sigma\left(W_i * [x_t, h_{t-1}] + b_i\right) \\ f_t &= \sigma\left(W_f * [x_t, h_{t-1}] + b_f\right) \\ o_t &= \sigma\left(W_o * [x_t, h_{t-1}] + b_o\right) \\ g_t &= \tanh\left(W_g * [x_t, h_{t-1}] + b_g\right) \\ c_t &= f_t \odot c_{t-1} + i_t \odot g_t \\ h_t &= o_t \odot \tanh\left(c_t\right), \end{aligned} \tag{2}$$

where c and h are the cell and hidden states as in traditional LSTMs, t is the corresponding timestep, x is the input data, $*$ the convolutional operation, $\odot$ the Hadamard product, and σ the sigmoid activation function. W_i, W_f, W_o, W_g and b_i, b_f, b_o, b_g are the learned weights and biases for the four convolutional operations of the ConvLSTM cell. The cell output is obtained from the last value of h.

Additionally, to account for the particularities of 360° content, and based on previous works [38], all the convolutional operations are replaced by spherical convolutions (Spherical ConvLSTM), a type of distorted convolutions that compute real pixel neighboring in spherical space. In the chosen implementation, the kernel is defined as the projection on the sphere of a small patch tangent to its surface. Thus, the kernel sampling pattern is distorted along with the image in the equirectangular projection. Following SphereNet's implementation, we employed spherical convolutions with a kernel size of 3×3. Therefore, the weights W_i, W_f, W_o and W_g in Eq. (2) have a size of $42 \times 36 \times 3 \times 3$ for the Spherical ConvLSTM of the encoder and $37 \times 1 \times 3 \times 3$ for that of the decoder.

3.2. Loss function

This model has been trained with a novel spherical weighted Kullback–Leibler Divergence (KLDiv) loss term, which refers to the KLDiv adapted to 360° content. The traditional Kullback–Leibler Divergence is a metric broadly used for saliency map comparisons, which measures the overall dissimilarity between two saliency maps that are considered as probability density functions. It can be computed as follows:

$$KLDiv\,(G, P) = \sum_{i,j} G_{i,j}\ \log\left(\epsilon + \frac{G_{i,j}}{\epsilon + P_{i,j}}\right), \tag{3}$$

where $P_{i,j}$, $G_{i,j}$, are the predicted and ground truth saliency values at pixel (i, j), and ϵ is a regularization constant that determines how much zero-valued predictions are penalized.

However, this metric does not account for the distortion introduced by the equirectangular projection, obtaining inadequate scores for this type of content. Two identical regions with equal areas on the spherical domain should obtain different scores depending on their location. Regions closer to the poles would produce larger projected areas in the 2D plane than those closer to the equator, resulting in higher KLDiv scores. To overcome this limitation and compensate for the distortion, Xu et al. [39] applied a spherical weighting by the solid angle to their Mean Squared Error (MSE) which presented the same problem. For an equirectangular image of shape n x m these weights can be computed as:

$$w_{i,j} = \Delta\varphi\ \frac{\sin(\Delta\theta j) + \sin(\Delta\theta(j+1))}{2\Delta\theta}, \tag{4}$$

where $\Delta\theta = \frac{\pi}{m}$; $\Delta\varphi = \frac{2\pi}{n}$; $i \in [0, n]$; $j \in [0, m]$.

Our proposed spherical weighted KLDiv loss employed to train the model applies this weighting to the traditional KLDiv, compensating for the distortion introduced by the equirectangular projection since the contribution of each pixel to the KLDiv is now proportional to its solid angle. Therefore, the proposed spherical weighted KLDiv loss is defined as follows:

$$\mathcal{L}_{KLDiv}\,(G, P) = \sum_{i,j} w_{i,j}\ G_{i,j}\ \log\left(\epsilon + \frac{G_{i,j}}{\epsilon + P_{i,j}}\right). \tag{5}$$

3.3. Datasets and training details

Our model was trained with the VR-EyeTracking dataset [3], whose videos present varied content (e.g., indoor scenes, outdoor activities, or sports games). They were down-sampled from 25 fps to 8 fps, and reshaped to a 320×240 resolution to reduce memory, computation, and processing requirements. Each video is then divided into sequences of 20 frames, corresponding to 2.5 s. From the entire dataset, we select the subset of 360° videos filmed with a static camera, since we use the optical flow as a measure of motion. Additionally, the subset of 144 videos is filtered by removing the videos whose optical flow provided by RAFT consisted only of noise (8% of the videos). The 360° videos

from the Sports-360 dataset [51] presenting a static camera were also used for an additional evaluation of the model and further comparisons to state-of-the-art works.

The ground truth saliency maps employed for training and evaluation were obtained from the eye fixations provided with the datasets using the method adopted by Sitzman et al. [2], employing a Gaussian filter with a standard deviation of 5° of visual angle, as proposed by Chao et al. [52]. The model has a size of 8,323 KB and was implemented within the Pytorch framework [53]. The training process took 15.65 h on a Nvidia RTX 2080 Ti with 11 GB of GDDR6 VRAM, using an Intel Xeon Gold G140 CPU at 3.7 GHz. We used the following hyperparameters for training: a stochastic gradient descent optimizer, a learning rate of 0.8, a batch size of 20 frames, and a total of 240 epochs. An average inference time of 49.923 ms (STD = 1.730 ms) was obtained after performing 3,000 model inferences with sequences of 20 frames on a GPU with the aforementioned specifications. Performing inference with several input image resolutions shows a linear increase of inference time with the number of pixels of the input image.

4. Results and evaluation

We conducted an in-depth analysis of the performance of our model for dynamic 360° saliency prediction. We present some results achieved with our proposed model (Section 4.2), and qualitative and quantitative comparisons with state-of-the-art works (Section 4.3). Furthermore, in Section 4.4 we offer detailed ablation studies that validate the effectiveness of the different elements included in our network architecture.

4.1. Saliency metrics

To assess the performance of the developed model, and provide a meaningful comparison with state-of-the-art works, we compute three different metrics commonly used in the literature for saliency maps comparison: Linear Correlation Coefficient (CC), Similarity Metric (SIM), and Kullback–Leibler Divergence (KLDiv). They are computed following the implementation proposed by Gutiérrez et al. [54] in which each pixel of a saliency map is weighted by the sine of its latitude, thus accounting for the distortion introduced by the equirectangular projection.

The SIM, CC, and KLDiv values presented as results in the comparisons represent the average mean and the average standard deviation of the measures obtained for all the videos. For each video, the mean and standard deviation are computed by evaluating all video frames with respect to their ground truth counterpart. However, how the ground truth is represented, whether the prediction includes dataset bias, whether the inputs are probabilistic, or whether spatial deviations exist between the prediction and ground truth can alter the scores [55]. Therefore, we provide three different metrics together with a qualitative analysis to draw conclusions and assess the true performance of the models.

4.2. Results

Figs. 1 and 3 show some sample results obtained with the proposed model in the VR-EyeTracking and Sports-360 datasets. For a set of videos, some consecutive RGB frames are shown together with their ground truth saliency and the model's prediction, blended with the frame. It can be seen how our proposed model is able to perform accurate saliency predictions, which are close to the ground truth, both focusing on small, yet relevant regions of the scene. In Fig. 3, it can also be appreciated the ability of our model to handle dynamic features, where although there are multiple salient elements (i.e., bicycles, people and cats), the network is able to detect and focus on the truly salient, thanks to the temporal information it has retained about what happened before in the 360° video. Please refer to the supplementary material for more qualitative results.[1]

Additionally, Table 1 provides a quantitative evaluation of the proposed model with the CC, SIM and KLDiv metrics obtained for the VR-EyeTracking and Sports-360 datasets, showing promising results with respect to state-of-the-art works.

4.3. Comparison with previous methods

We compared our model against three state-of-the-art works whose implementation is publicly available: one spherical DNN designed for saliency prediction in 360° images (i.e., Martin et al.'s [40]) to serve as a baseline, and to show that static approaches are not suitable for dynamic content, and two state-of-the-art models specialized on 360° videos, ATSal [43] and CP-360 [42]. To assess the models' performance, the metrics CC, SIM, and KLDiv (Section 4.1) were computed over the saliency predictions of the models for the VR-EyeTracking test split and the Sports-360 dataset.

Table 1 shows that the proposed model outperforms previous approaches in all computed metrics. Martin et al.'s model was trained on a dataset of static 360° images, which hinders its ability to generalize to dynamic content, focusing on more low-level features such as contrast or edges. Both ATSal and CP-360 models were trained on dynamic content, and have a small LSTM-based module on their bottleneck to account for temporal relations. However, they neglect the fact that temporal information may be of importance even when extracting and encoding features, and are therefore limited. The proposed model, nevertheless, is able to handle temporal relations in all stages yielding more accurate results.

Additionally, Fig. 4 shows a qualitative comparison between the models on a small subset of 360° sequences extracted from the VR-EyeTracking and Sports-360 datasets. In these sampled sequences we can see how ATSal fails to identify the truly salient regions, CP-360 tends to overestimate the saliency, and Martin et al.'s produces saliency maps with a strong equator bias, typical of the static content but not present in the dynamic case. Our model offers the most plausible saliency predictions, correctly identifying the regions where humans tend to direct their attention.

4.4. Ablation studies

To endorse the effectiveness of the decisions taken through the design stage of the model's architecture, we performed five ablation studies. These studies analyze the influence of the input data resolution, the use of spherical convolutions, the loss function employed, the inclusion of optical flow, and the advantages reported by the Spherical ConvLSTMs. The results obtained for each experiment of the ablation studies are shown in Table 2, where the values represent the average mean and standard deviation obtained for the test split of the VR-EyeTracking dataset.

4.4.1. Input resolution

Firstly, we analyzed the influence of the input resolution of the 360° videos on the ability of the network to predict accurate saliency maps. We trained our model with two different setups: with the 360° videos at a resolution of 208 × 106, and with a higher resolution of 320x240. The quantitative results obtained can be seen in the first two rows of Table 2, in which the

[1] https://graphics.unizar.es/projects/SST-Sal_2022.

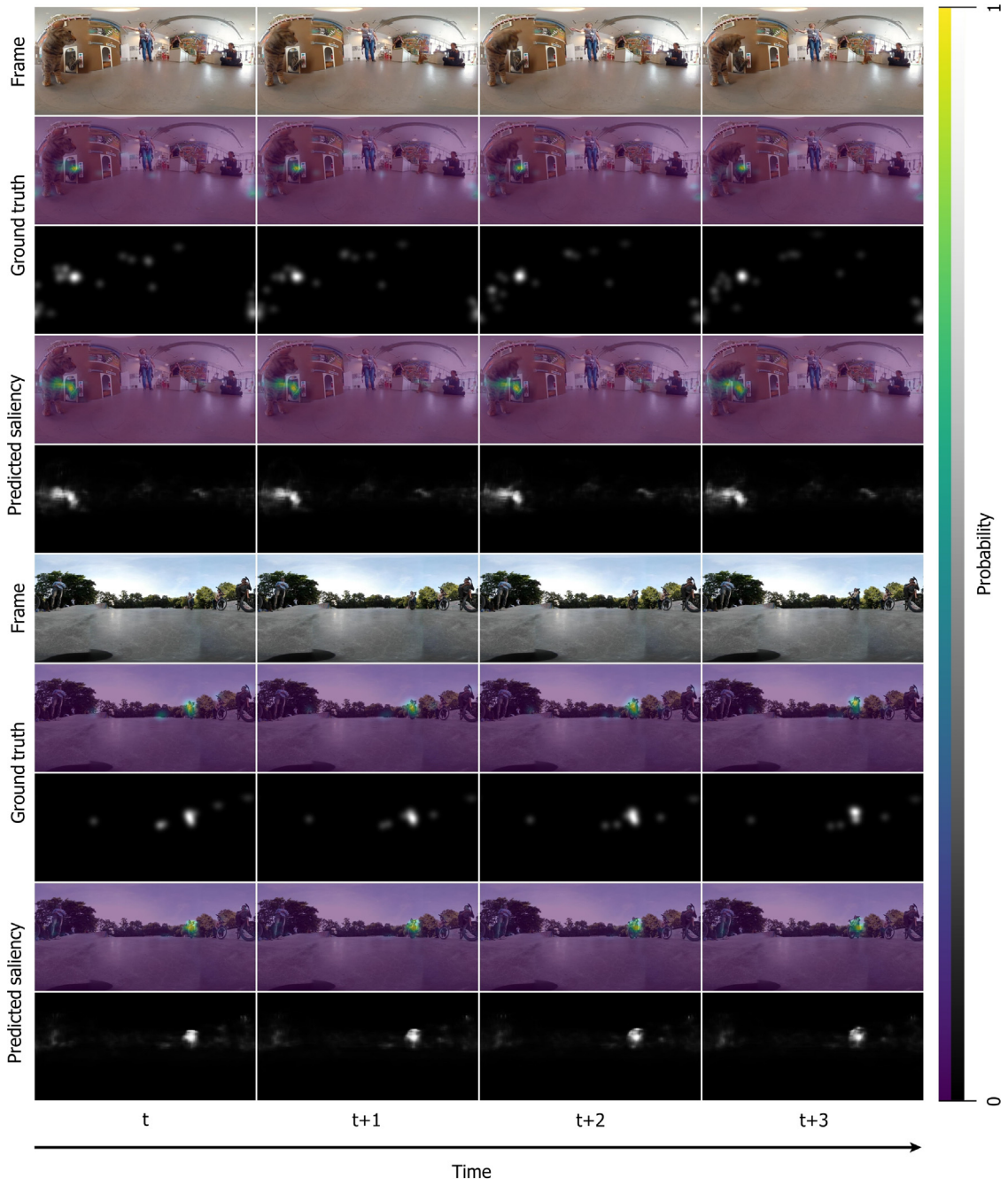

Fig. 3. Results obtained with the proposed model for two sequences of the VR-EyeTracking (top) and Sports-360 (bottom) datasets. The horizontal axis represents time. The vertical axis shows the frames, the ground-truth saliency, and the predicted saliency of the sequences. Saliency is represented both in greyscale, and as a heat map blended with the frame's image, where yellowish colors correspond to more salient areas. Note that the proposed model performs accurate predictions similar to the ground truth, both focusing in small, yet relevant regions of the scene.

mean score of the three metrics drops when using the lower resolution layout. This could indicate that the reduction of the frames' resolution hindered the training process, probably due to a significant loss of information.

4.4.2. Spherical convolutions

The present ablation study justifies in terms of the model's performance the inclusion of the spherical convolutions on the ConvLSTM architecture. For its evaluation, the proposed model was modified, replacing the spherical convolutions implemented in the Spherical ConvLSTMs of both encoder and decoder by traditional convolutions. Then, the non-spherical model is trained under the same conditions as the original spherical one. The results in Table 2 show an improvement in the model's performance when spherical convolutions are used. This evidences that the use of architectures that account for the peculiarities of the equirectangular representation is an advantageous approach when dealing with 360° content.

4.4.3. Alternative loss functions

The current spherical weighted KLDiv loss function used to train the model was compared against two alternative configurations. A traditional non-spherical KLDiv (Eq.), in which all pixels have equal weight in the metric no matter where they are located, and a spherical sampled KLDiv, which presents a different strategy to account for the distortion introduced by the

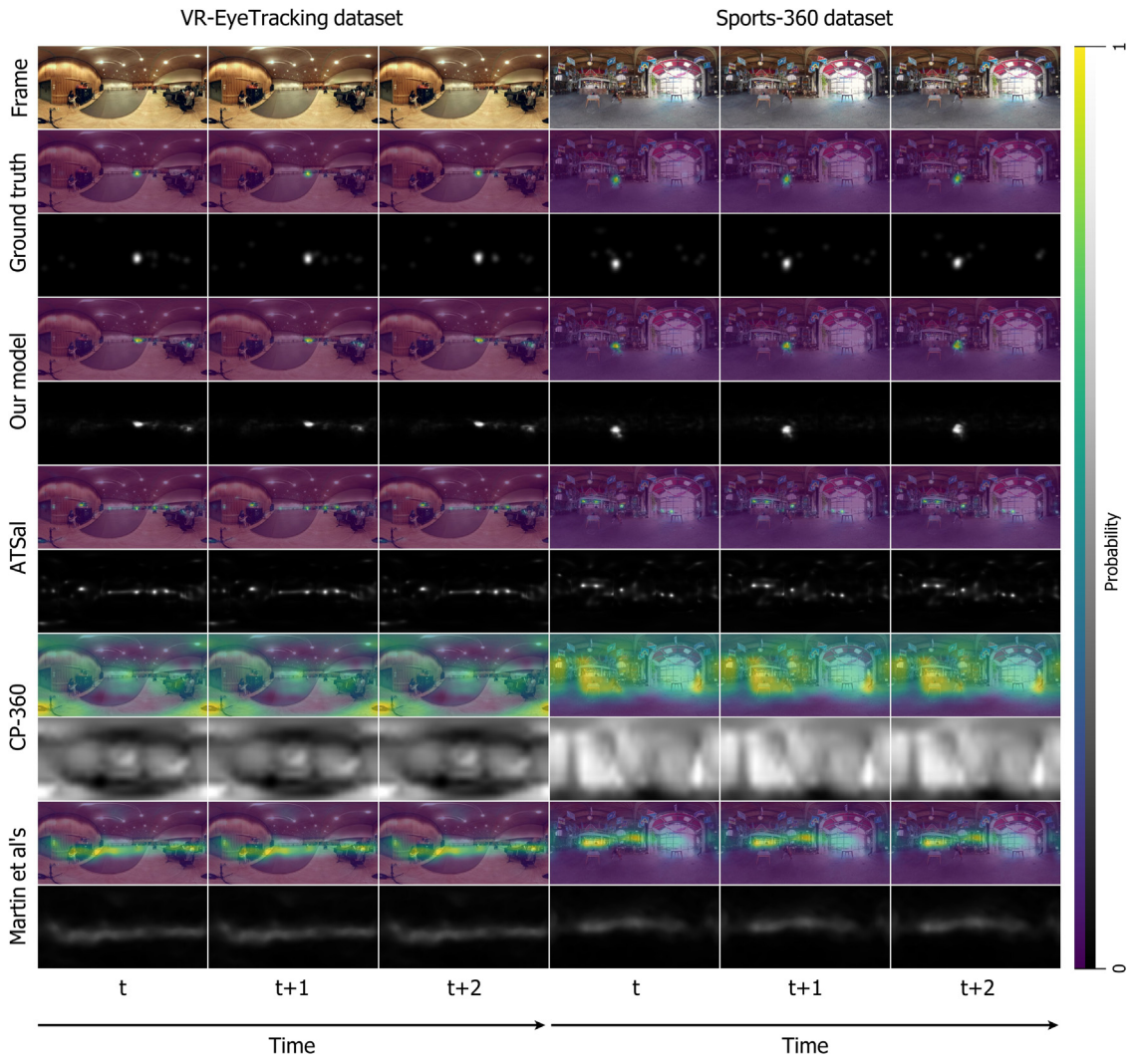

Fig. 4. Qualitative comparison of the proposed model against ATSal [43], CP-360 [42], and Martin et al.'s [40]. The horizontal axis represents time. The vertical axis shows, from top to bottom: The RGB frame, the ground-truth saliency, the proposed model's prediction, ATSal, CP-360, and Martin et al.'s saliency predictions. Saliency is shown both with a heat map blended with the frame's image (yellowish colors correspond to high salient areas) and with the saliency maps (black indicates zero probability of being observed). Note that the proposed model outperforms the state-of-the-art works since its saliency prediction is closer to the ground truth. Martin et al.'s, as a saliency prediction method for 360° images, tends to predict exploration saliency maps typical of this type of content. Both ATSal and CP-360 models were trained on *dynamic* 360° content, but either fail to identify the truly salient elements of the 360° video, or predict a high probability for most of the frame's regions, which does not resemble the ground-truth behavior. ∗ ATSal was trained with the VR-EyeTracking dataset; its results are included for completeness.

Table 1
Quantitative comparisons of the proposed model against ATSal [43], CP-360 [42], and Martin et al.'s [40], with both VR-EyeTracking and Sports-360 datasets. Arrows indicate whether higher or lower is better, and boldface highlights the best result for each metric. The values represent the mean score among the different videos in the dataset for each metric, and in brackets is shown the averaged standard deviation. ∗ ATSal was trained with the VR-EyeTracking dataset; its results are included for completeness.

	VR-EyeTracking dataset			Sports-360 dataset		
	CC ↑	SIM ↑	KLDiv ↓	CC ↑	SIM ↑	KLDiv ↓
ATSal*	0.298 (0.087)	0.216 (0.041)	9.858 (1.000)	0.246 (0.090)	0.184 (0.050)	10.552 (1.221)
CP-360	0.229 (0.049)	0.148 (0.025)	10.665 (0.556)	0.228 (0.055)	0.135 (0.031)	11.753 (0.731)
Martin en al.'s	0.138 (0.054)	0.152 (0.033)	8.610 (0.941)	0.240 (0.078)	0.183 (0.049)	11.191 (1.176)
Our Model	**0.500 (0.123)**	**0.338 (0.058)**	**7.371 (1.218)**	**0.439 (0.143)**	**0.284 (0.070)**	**8.610 (1.591)**

equirectangular projection. Instead of weighting each pixel by its solid angle, the equirectangular image's pixels are sampled following a uniform cosine sampling distribution. Therefore, the equator of the image will have more representation than the poles on the final value of the loss function, since a greater number of pixels belonging to this region will be used to compute the metric.

Table 2
Ablation Study: CC, SIM and KLDiv scores obtained with the different configurations of the ablation studies for the VR-EyeTracking dataset test split. The values represent the mean score among the different videos in the dataset for each metric, and in brackets is shown the average standard deviation. Check marks indicate which elements are used in the ablation study, and boldface highlights the best result for each metric. Please refer to Sections 4.4.1 to 4.4.5 for further details.

Resolution 208 × 106	Resolution 320 × 240	Spherical convLSTM	Optical flow	Weighted KLDiv	Sampled KLDiv	Traditional KLDiv	CC ↑	SIM ↑	KLDiv ↓
✓	–	✓	✓	✓	–	–	0.413 (0.113)	0.287 (0.047)	8.458 (1.066)
–	✓	✓	✓	✓	–	–	**0.500 (0.123)**	**0.338 (0.058)**	**7.372 (1.218)**
–	✓	–	✓	✓	–	–	0.438 (0.136)	0.315 (0.067)	7.602 (1.467)
–	✓	✓	✓	–	–	✓	0.467 (0.121)	0.324 (0.059)	7.645 (1.286)
–	✓	✓	✓	–	✓	–	0.474 (0.099)	0.310 (0.051)	8.066 (1.133)
–	✓	✓	–	✓	–	–	0.480 (0.112)	0.317 (0.053)	7.924 (1.166)
–	✓	–	✓	✓	–	–	0.344 (0.122)	0.213 (0.043)	10.097 (0.938)

The results showed in Table 2 suggest that the most suitable loss is the weighted KLDiv, as it achieves better accuracy for all the metrics. It was expected that the sampled KLDiv loss function, which accounts for the equirectangular projection, would report better performance than the traditional KLDiv loss. However, the sampled KLDiv loss presents worse results for SIM and CC. This could be either because of the poor performance of the randomness of the samples, or the fact that most of the relevant information of the videos is in the less distorted part, leading the traditional KLDiv loss to reach a reasonably accurate performance.

4.4.4. Inclusion of optical flow estimation

The motivation for providing an estimate of the optical flow between consecutive frames is to help the network to identify salient features, since moving elements tend to capture human visual attention. Therefore, the aim of this ablation study is to analyze the effect that providing the network with this information has on its performance. For this purpose, we trained our model depriving it of the optical flow estimations. The results obtained with this configuration (see Table 2) show an improvement in performance when including the optical flow. Without it, the model may overestimate the saliency of some static features, which while salient in static content, are unattractive in dynamic content due to the existence of motion and action.

4.4.5. Influence of spherical convlstms

The estimation of the optical flow provides the network with temporal information about the previous frame. The present study aims to determine if the temporal information provided by the optical flow estimation is sufficient, or if ConvLSTMs need to be included to infer the temporal relationships between frames. For this purpose, a structure without Spherical ConvLSTMs is trained. The alternative model consists of an encoder and decoder formed only by spherical convolutional layers. The encoder takes as input the concatenation of the RGB frame and RGB optical flow and applies a spherical convolution and a spherical Max Pool (SphereNet's [38] implementation). Then, the decoder returns the predicted saliency map by processing the encoded feature vector with another spherical convolution and an Up-Sampling layer.

Table 2 shows a lower performance when using only optical flow estimations than when using only Spherical ConvLSTMs, indicating that the temporal information provided by the optical flow is not enough, and ConvLSTMs are needed to infer temporal relationships between frames. However, the poor accuracy reported may be due to the simplicity of the structure, which is built with only two spherical convolutional layers, whereas deeper architectures are usually employed in DNNs for image processing. Nevertheless, this simple architecture provides a fair comparison with our model, since with the same number of convolutions and equal encoded feature vector's size, it achieves better saliency predictions. Although the temporal information provided by the optical flow is not sufficient, it does have value. Therefore, our model takes advantage of using the optical flow jointly with the Spherical ConvLSTMs, reaching the best performance of all the presented alternatives.

5. Conclusions

We propose a novel saliency prediction model for dynamic 360° content that is able to resemble human visual behavior, outperforming previous state-of-the-art works. Our network is built over Spherical ConvLSTM cells, and it has been specifically tailored to the peculiarities of 360° content with the inclusion of spherical convolutions and a novel spherical KLDiv loss function. Additionally, our model benefits from the information provided by optical flow estimations between consecutive frames, which enhances the final saliency prediction.

Our work suggests that dynamic content requires models that account for temporal information. Architectures specifically designed for videos surpass saliency prediction models designed for static images, since the awareness of the temporal relationships between frames seems to play a major role. Furthermore, the availability of this temporal information at the encoding and decoding stages has proved to be an advantageous approach, showing outstanding results with respect to previous state-of-the-art works. To the best of our knowledge, our model has been the first to account for this temporal information in feature extraction time.

6. Limitations and future work

Our model for saliency prediction in 360° videos was trained with a selection of videos from the VR-EyeTracking dataset filmed without camera movement. Static videos are a common scenario in 360° video acquisition due to the difficulty of generating compelling content. Wide-angle cameras that are capable of filming high-quality 360° videos are often made up of strategically placed, narrow-angle cameras. This implies a considerable size

and weight, limiting the camera's manoeuvrability. Stitching algorithms are also needed to unify the independent images, which are often problematic when working with moving cameras. Thus, saliency prediction on 360° videos recorded with a dynamic camera remains an interesting future avenue.

Like motion, the depth of the elements is a determining factor in their saliency [56,57]. Therefore, feeding the network with depth estimations could provide a better understanding of the scene, enhancing the predicted saliency for the 360° videos.

Moreover, dynamic 360° environments are not only visual experiences, but they usually present multimodal stimuli that may affect how observers perceive and explore the scenes [58]. Like directional acoustic cues, which can greatly alter visual perception [59] and are currently being overlooked. Therefore, the study of the inclusion of mono or spatial audio together with the 360° video would be an interesting line for future work.

We believe that this work is a timely step towards understanding and modeling human visual behavior in dynamic 360° environments.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work has received funding from the European Union's Horizon 2020 research and innovation programme (ERC project CHAMELEON, Grant No 682080, and Marie Sklodowska-Curie project Grant No 956585). This project was also supported by a 2020 Leonardo Grant for Researchers and Cultural Creators, BBVA Foundation (the BBVA Foundation accepts no responsibility for the opinions, statements, and contents included in the project and/or the results thereof, which are entirely the responsibility of the authors). This work has also received funding from Spain's Agencia Estatal de Investigación, Spain (project PID2019-105004GB-I00). Additionally, Daniel Martin and Edurne Bernal were supported by a Gobierno de Aragon predoctoral grant (2020–2024 and 2021–2025, respectively). Edurne Bernal also received a grant from the Instituto Universitario de Investigacion en Ingenieria de Aragon (I3A).

References

[1] Martin D, Serrano A, Bergman AW, Wetzstein G, Masia B. ScanGAN360: A generative model of realistic scanpaths for 360° images. IEEE Trans Vis Comput Graphics 2022;28(5):2003–13.
[2] Sitzmann V, Serrano A, Pavel A, Agrawala M, Gutierrez D, Masia B, et al. Saliency in VR: How do people explore virtual environments? IEEE Trans Vis Comput Graphics 2018;24(4):1633–42.
[3] Xu Y, Dong Y, Wu J, Sun Z, Shi Z, Yu J, et al. Gaze prediction in dynamic 360° immersive videos. In: IEEE/CVF Conference on computer vision and pattern recognition. 2018, p. 5333–42.
[4] Arabadzhiyska E, Tursun OT, Myszkowski K, Seidel H-P, Didyk P. Saccade landing position prediction for gaze-contingent rendering. ACM Trans Graph (Proc. SIGGRAPH) 2017;36(4).
[5] Hu Z, Zhang C, Li S, Wang G, Manocha D. SGaze: A data-driven eye-head coordination model for realtime gaze prediction. IEEE Trans Vis Comput Graphics 2019;25(5):2002–10.
[6] Hu Z, Li S, Zhang C, Yi K, Wang G, Manocha D. DGaze: CNN-based gaze prediction in dynamic scenes. IEEE Trans Vis Comput Graphics 2020;26(5):1902–11.
[7] Luz G, Ascenso J, Brites C, Pereira F. Saliency-driven omnidirectional imaging adaptive coding: Modeling and assessment. In: IEEE International workshop on multimedia signal processing. 2017, p. 1–6.
[8] Serrano A, Sitzmann V, Ruiz-Borau J, Wetzstein G, Gutierrez D, Masia B. Movie editing and cognitive event segmentation in virtual reality video. ACM Trans Graph (Proc. SIGGRAPH) 2017;36(4).
[9] Cao Y, Lau R, Chan AB. Look over here: Attention-directing composition of manga elements. ACM Trans Graph (Proc. SIGGRAPH) 2014;33.
[10] Itti L, Koch C, Niebur E. A model of saliency-based visual attention for rapid scene analysis. IEEE Trans Pattern Anal Mach Intell 1998;20(11):1254–9.
[11] Itti L, Koch C. A saliency-based search mechanism for overt and covert shifts of visual attention. Vis Res 2000;40:1489–506.
[12] Erdem E, Erdem A. Visual saliency estimation by nonlinearly integrating features using region covariances.. J Vis 2013;13 4:11.
[13] Bruce N, Tsotsos J. Saliency based on information maximization. Adv Neural Inf Proc Syst 2006;18.
[14] Harel J, Koch C, Perona P. Graph-based visual saliency. Adv Neural Inf Process Syst 2006;19.
[15] Borji A. Boosting bottom-up and top-down visual features for saliency estimation. In: IEEE Conference on computer vision and pattern recognition. 2012, p. 438–45.
[16] Judd T, Ehinger KA, Durand F, Torralba. A. Learning to predict where humans look. In: IEEE International conference on computer vision. 2009, p. 2106–13.
[17] Lee S-H, Kim J-H, Choi KP, Sim J-Y, Kim C-S. Video saliency detection based on spatiotemporal feature learning.. In: IEEE International conference on image processing. 2014, p. 1120–4.
[18] Kruthiventi SSS, Ayush K, Babu RV. DeepFix: A fully convolutional neural network for predicting human eye fixations. IEEE Trans Image Process 2017;26(9):4446–56.
[19] Liu N, Han J, Zhang D, Wen S, Liu T. Predicting eye fixations using convolutional neural networks. In: IEEE Conference on computer vision and pattern recognition. 2015, p. 362–70.
[20] Kümmerer M, Theis L, Bethge M. Deep gaze I: Boosting saliency prediction with feature maps trained on ImageNet. 2015, arXiv.
[21] Pan J, Sayrol E, Giro-I-Nieto X, McGuinness K, O'Connor NE. Shallow and deep convolutional networks for saliency prediction. In: IEEE Conference on computer vision and pattern recognition. 2016, p. 598–606.
[22] Pan J, Sayrol E, Nieto XG-i, Ferrer CC, Torres J, McGuinness K, et al. Salgan: Visual saliency prediction with adversarial networks. In: CVPR Scene understanding workshop. 2017.
[23] Cornia M, Baraldi L, Serra G, Cucchiara R. Predicting human eye fixations via an LSTM-based saliency attentive model. IEEE Trans Image Process 2018;27(10):5142–54.
[24] Liu N, Han J. A deep spatial contextual long-term recurrent convolutional network for saliency detection. IEEE Trans Image Process 2018;27(7):3264–74.
[25] Martin D, Gutierrez D, Masia B. A probabilistic time-evolving approach to scanpath prediction. 2022, arXiv.
[26] Bak C, Erdem A, Erdem E. Two-stream convolutional networks for dynamic saliency prediction. 2016, arXiv.
[27] Bak C, Kocak A, Erdem E, Erdem A. Spatio-temporal saliency networks for dynamic saliency prediction. IEEE Trans Multimed 2018;20(7):1688–98.
[28] 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. 2018, p. 602–17.
[29] Zhu Y, Zhai G, Min X. The prediction of head and eye movement for 360 degree images. Signal Process, Image Commun 2018;69:15–25.
[30] Battisti F, Baldoni S, Brizzi M, Carli M. A feature-based approach for saliency estimation of omni-directional images. Signal Process, Image Commun 2018;69:53–9.
[31] Fang Y, Zhang X, Imamoglu N. A novel superpixel-based saliency detection model for 360-degree images. Signal Process, Image Commun 2018;69:1–7.
[32] Lebreton P, Raake A. GBVS360, BMS360, ProSal: Extending existing saliency prediction models from 2D to omnidirectional images. Signal Process, Image Commun 2018;69:69–78.
[33] Luz G, Ascenso J, Brites C, Pereira F. Saliency-driven omnidirectional imaging adaptive coding: Modeling and assessment. In: IEEE international workshop on multimedia signal processing. 2017, p. 1–6.
[34] Startsev M, Dorr M. 360-Aware saliency estimation with conventional image saliency predictors. Signal Process, Image Commun 2018;69:43–52.
[35] Assens M, Giro-i Nieto X, McGuinness K, O'Connor NE. SaltiNet: Scanpath prediction on 360 degree images using saliency volumes.. In: IEEE International conference on computer vision workshops. 2017, p. 2331–8.
[36] Monroy R, Lutz S, Chalasani T, Smolic A. SalNet360: Saliency maps for omni-directional images with CNN. Signal Process, Image Commun 2018;69:26–34.
[37] Lv H, Yang Q, Li C, Dai W, Zou J, Xiong H. SalGCN: Saliency prediction for 360-degree images based on spherical graph convolutional networks. In: MM 2020 - Proceedings of the 28th ACM International conference on multimedia. 2020, p. 682–90.
[38] Coors B, Condurache AP, Geiger A. SphereNet: Learning spherical representations for detection and classification in omnidirectional images.. In: Ferrari V, Hebert M, Sminchisescu C, Weiss Y, editors. Computer vision – ECCV 2018. Cham: Springer International Publishing; 2018, p. 525–41.

[39] Xu Y, Zhang Z, Gao S. Spherical DNNs and their applications in 360° images and videos. IEEE Trans Pattern Anal Mach Intell 2021.
[40] Martin D, Serrano A, Masia B. Panoramic convolutions for 360° single-image saliency prediction. In: CVPR Workshop on computer vision for augmented and virtual reality. 2020.
[41] Nguyen A, Yan Z, Nahrstedt K. Your attention is unique: Detecting 360-degree video saliency in head-mounted display for head movement prediction.. In: MM 2018 - Proceedings of the 2018 ACM multimedia conference. 2018, p. 1190–8.
[42] Cheng H-T, Chao C-H, Dong J-D, Wen H-K, Liu T-L, Sun M. Cube padding for weakly-supervised saliency prediction in 360° videos. In: IEEE/CVF Conference on computer vision and pattern recognition. 2018, p. 1420–9.
[43] Dahou Y, Tliba M, McGuinness K, O'Connor N. ATSal: An attention based architecture for saliency prediction in 360 videos. Lecture Notes in Comput Sci 2020;12663 LNCS:305–20.
[44] Linardos P, Mohedano E, Nieto JJ, O'Connor NE, i Nieto XG, McGuinness K. Simple vs complex temporal recurrences for video saliency prediction. In: BMVC. 2019.
[45] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. In: MICCAI. 2015.
[46] Pratt J, Radulescu PV, Guo RM, Abrams RA. It's alive!: Animate motion captures visual attention. Psychol Sci 2010;21(11):1724–30, PMID: 20974713.
[47] Meyerhoff H, Schwan S, Huff M. Perceptual animacy: Visual search for chasing objects among distractors. J Exp Psychol Hum Percept Perform 2013;40.
[48] Teed Z, Deng J. Raft: Recurrent all-pairs field transforms for optical flow. In: European conference on computer vision. Springer; 2020, p. 402–19.
[49] Butler DJ, Wulff J, Stanley GB, Black MJ. A naturalistic open source movie for optical flow evaluation. In: European conf. on computer vision. ECCV, LNCS, vol. 7577, 2012, p. 611–25, Part IV.
[50] Shi X, Chen Z, Wang H, Yeung D-Y, Wong W-k, Woo W-c. Convolutional LSTM network: A machine learning approach for precipitation nowcasting.. In: Proceedings of the 28th International conference on neural information processing systems - volume 1. NIPS'15, Cambridge, MA, USA: MIT Press; 2015, p. 802–10.
[51] Zhang Z, Xu Y, Yu J, Gao S. Saliency detection in 360 videos. In: Proceedings of the european conference on computer vision. 2018, p. 488–503.
[52] Chao F-Y, Ozcinar C, Wang C, Zerman E, Zhang L, Hamidouche W, et al. Audio-visual perception of omnidirectional video for virtual reality applications.. In: IEEE International conference on multimedia expo workshops. 2020, p. 1–6.
[53] Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: An imperative style, high-performance deep learning library. In: Adv neural inf proc syst, vol. 32. Curran Associates, Inc.; 2019, p. 8024–35.
[54] Gutiérrez J, David E, Rai Y, Callet PL. Toolbox and dataset for the development of saliency and scanpath models for omnidirectional 360° still images. Signal Process, Image Commun 2018;69:35–42.
[55] Bylinskii Z, Judd T, Oliva A, Torralba A, Durand F. What do different evaluation metrics tell us about saliency models? IEEE Trans Pattern Anal Mach Intell 2019;41(3):740–57.
[56] Laco M, Benesova W. Depth in the visual attention modelling from the egocentric perspective of view. In: Eleventh international conference on machine vision, vol. 11041. SPIE; 2019, p. 329–39.
[57] Desingh K, Krishna KM, Rajan D, Jawahar CV. Depth really matters: Improving Visual Salient Region detection with depth.. In: BMVC. 2013.
[58] Martin D, Malpica S, Gutierrez D, Masia B, Serrano A. Multimodality in VR: A survey. ACM Comput Surv 2022.
[59] Masia B, Camon J, Gutierrez D, Serrano A. Influence of directional sound cues on users' exploration across 360° movie cuts. IEEE Comput Graph Appl 2021;41(4):64–75.