Distortion-Induced Saliency Shifts in Video

Xinbo Wu, Jianxun Lou, Zhengyan Dong, Fan Zhang, Paul Rosin and Hantao Liu

***Abstract*—Visual saliency modelling is of fundamental importance in modern video processing and its applications. Our previous eye-tracking study revealed that signal distortions caused by editing, compression, or transmission alter gaze patterns and consequently induce saliency shifts in both spatial and temporal domains. Saliency shifts provide crucial insights into viewers' behavioural responses to video distortions, facilitating the perception-based optimisation of video algorithms. However, the spatio-temporal saliency shifts and their measurable effects on perception related applications remain largely unexplored. In this paper, we first investigate the measurement of distortion-induced saliency shifts (DSS) in videos and analyse DSS behaviours as functions of video content, time order and critical distortion disruption. Second, based on our findings, we construct three vision models to quantitatively simulate distinct DSS behaviours and integrate them into a comprehensive DSS behaviour model. Finally, we demonstrate that the computational DSS model can enhance emerging video technologies.**

***Index Terms*—Visual distortion, gaze, perception, eye-tracking, saliency shifts**

I. Introduction

THe use of digital videos in our daily lives has grown significantly over the past few decades. This is largely attributed to the rapid development of internet technologies and mobile devices, which enable easy access to video content through various digital platforms such as social media and online streaming services. Globally, consumer internet traffic is already dominated by internet video traffic, with more than one billion hours of video being watched every day [1].

Video signals are inevitably vulnerable to distortions caused by acquisition, compression, transmission, and display. These distortions not only affect consumers' visual experiences but also lead to misinterpretation of video content for many visual tasks [2]. It is critical to understand human perception of video distortions and use the knowledge to develop advanced technologies for video processing and its related applications, such as video compression, video enhancement, and video quality assessment. However, challenges remain in reliably capturing the way humans perceive distortions in concurrence with video content within a complex spatio-temporal context. To push forward research in this field, further effort is needed to better understand how the human visual system (HVS) responds to the combination of natural content and its distortions and then develop models to integrate these behavioural responses in video algorithms.

A significant research trend in video processing is to exploit visual saliency – an important mechanism of the HVS [3]. Visual saliency represents the HVS's ability to select and prioritise the most relevant information from a visual scene, which plays a significant role in perception and decision-making regarding visual content [4], [5]. This inherent feature of the HVS allows individuals to identify areas of interest and importance, facilitating efficient interaction with complex visual environments. Many video processing algorithms have taken advantage of computational saliency, integrating modelled saliency into various applications including compression, noise reduction, and quality assessment [6]. In general, traditional video algorithms that rely on handcrafted visual features often incorporate a dedicated computed saliency to enhance their accuracy and efficacy [7]. For deep learning-based video algorithms, saliency is integrated directly into the deep neural networks to augment the prediction task. By embedding saliency mechanisms, both traditional and deep learning-based algorithms inherently prioritise significant regions in the visual data, improving their overall performance. For example, in video compression, saliency models are utilised to allocate more resources to visually important areas while compressing less critical regions more aggressively [8]. By focusing on the regions that viewers are most likely to notice, saliency-aware compression algorithms ensure that the most critical parts of the video retain high quality, even at lower bit-rates. In noise reduction, visual saliency is used prioritise the preservation of salient details while effectively suppressing background noise [9]. This saliency-guided selective noise reduction process can achieve a better balance between noise suppression and detail preservation, resulting in perceptually optimised videos. Visual quality assessment has been a longstanding area of research interest [2], [10]–[12]. Within this field, video quality assessment (VQA) plays a critical role, with visual saliency serving as a key factor influencing perceived video quality [13]. A comprehensive survey is provided in [10] on perceptual image quality assessment (IQA), covering fundamental concepts, traditional models, and recent advancements driven by deep learning. The work in [11] presents an extensive overview of screen content quality assessment (SCQA), focusing on unique characteristics such as text, graphics, and mixed content types in screen-based visuals. The survey in [2] reviews state-of-the-art methods in perceptual video quality assessment (VQA), addressing how temporal dynamics, motion, and video distortions affect perceived quality. VQA algorithms often treat all regions of a video equally, which can lead to inaccurate evaluations of perceptual quality. Saliency-based VQA algorithms, however, focus on the regions that are most likely to draw viewers' attention, providing a more accurate measure of video quality as perceived by human observers. In [14], visual saliency is

Xinbo Wu, Jianxun Lou, Zhengyan Dong, Paul Rosin, and Hantao Liu are with the School of Computer Science and Informatics, Cardiff University, CF24 4AG Cardiff, United Kingdom.

Jianxun Lou is also with the School of Computer Science, Northeast Electric Power University, Jilin, China.

Fan Zhang is with the School of Computer Science, University of Bristol, BS8 1UB Bristol, United Kingdom.

Corresponding author: Jianxun Lou (jianxunlou@outlook.com)

modelled in the contrast sensitivity function (CSF), which is then integrated into a wavelet-based distortion visibility measure to build a foveated VQA model. In [15], a quality-aware visual attention module is established to obtain saliency-guided representations for an end-to-end blind VQA model.

The literature has shown that integrating visual saliency into video compression, noise reduction, and quality assessment algorithms significantly enhances their performance. However, the approaches taken so far rely on minimal modelling assumptions of the HVS – saliency is treated as a simplified weighting function for spatial and/or temporal distortions. This method overlooks the complex and dynamic nature of visual saliency. Consequently, it remains largely unexplored how saliency contributes to the overall performance of a video algorithm. It is of fundamental importance to have a better understanding of the underlying interactive mechanisms between saliency, natural content, and distortion. The foundation of saliency modelling is eye movements of human viewers, which carry critical information on perceptual-cognitive behaviour of humans in watching videos. A recent eye-tracking study [6] has taken a rigorous approach to prove that there is a significant difference in saliency between reference scenes (i.e., original and pristine video content) and distorted scenes (i.e., video content with visible distortions). Eye movements reflect visual attention through both a bottom-up, saliency-driven, and task-independent process, as well as a top-down, volition-controlled, and task-dependent process [16], [17]. Saliency, representing bottom-up and task-independent visual attention, plays a crucial role in understanding spontaneous gaze patterns. Free-viewing eye-tracking is widely regarded as the "gold standard" for measuring saliency due to its ecological validity and avoidance of task-induced biases [18]. Saliency maps, derived from robust eye-tracking protocols with sufficient participants, provide a consistent and generalizable representation of bottom-up visual attention at the population level [19]. Previous research has demonstrated their reliability across different observer groups and experimental settings [20]. It should be noted that while eye movement patterns observed during a video quality rating task reflect a combination of both bottom-up (saliency) and top-down attention, the saliency component aligns with the eye movement patterns observed in a free-viewing task, highlighting the robustness and applicability of saliency in diverse visual computing contexts [6]. In this paper, we conduct systematic analyses of observers' gaze behaviour in terms of saliency shifts induced by distortions in video. We investigate how the degree of saliency shifts is affected by three significant video properties, i.e., video content, passage of time, and critical distortion disruption. These behavioural responses are then used to derive vision models to characterise the perception of video distortions and consequently to enhance video algorithms.

The contributions of this work are as follows:

- First, leveraging large-scale video eye-tracking data, we conduct a thorough statistical analysis to reveal human behavioural responses to distortions in video. We introduce the concept of distortion-induced saliency shifts (DSS) and define it as a function of video content, time order and critical distortion disruption.
- Second, we develop three vision models to quantitatively simulate DSS behaviours. The first model examines the impact of video content on DSS; the second model analyses the effect of passage of time on DSS; and the third model identifies the frames that cause significant DSS in a video.
- Finally, we propose a novel, generic DSS behaviour fusion model that integrates the three vision models, and demonstrate the model's effectiveness in enhancing video algorithms.

II. Related work

In the literature, eye-tracking studies have been undertaken to provide a fundamental understanding of the phenomenon of distortion-induced saliency shifts (DSS) – the difference in saliency between pristine visual content and its distorted format. An eye-tracking experiment was carried out in [21] to study the impact of visual distortions on the fixation patterns. The saliency maps of an undistorted image and its distorted version were visualised and compared via visual inspection. It is observed that distortions such as white noise, blurring and compression artefacts can significantly alter the saliency of the undistorted image. A further eye-tracking study [22] was carried out to specifically investigate the impact of JPEG compression artefacts on fixation patterns. It is found that JPEG compression artefacts when introduced to a pristine image can attract attention and consequently alter the saliency of the image; and that the degree of saliency alteration is related to the strength of artefacts. In [23], an eye-tracking experiment was performed to analyse the deviation in saliency from natural/undistorted scene saliency as a consequence of introducing visual distortions including Gaussian blur, white noise and JEPG compression. The study revealed that the difference between the natural/undistorted scene saliency and deviated saliency caused by distortions is significant, and that the lower the quality of the distorted image the higher the deviation is from the natural/undistorted scene saliency. It's worth noting that these studies are primarily focused on still images, and research on DSS in videos is still relatively limited. Some attempts have been made in the literature e.g., an eye-tracking experiment was performed in [24] to understand how people watch a video sequence. The set of stimuli included 10 original video sequences and 50 impaired video sequences (i.e., five levels of impairments obtained by H.264 video compression). The gaze patterns/allocations of stimuli were compared to measure the impact of distortions on the visual attention deployment. In summary, existing eye-tracking studies have demonstrated that visual distortions can cause saliency to shift from its original places in the pristine/undistorted content, and such phenomenon of distortion-induced saliency shifts (DSS) provides insights into how saliency plays a role in visual perception. However, the above-mentioned eye-tracking studies exhibit few limitations: (1) a limited number of subjects and/or a small degree of stimulus variability are used in the experiments, which limits the generality of the findings; (2) eye-tracking data is often biased/contaminated due to the involvement of stimulus repetition (i.e., carry-over effects [6]) – each observer is asked to view the same natural scene content (rendered with multiple variations by adding distortions) repeatedly – hence cannot be used as the reliable ground truth to study the distortion-induced saliency shifts (DSS); and (3) in some experiments eye-tracking data is collected under task-specific conditions, where gaze behaviour is primarily driven by the tasks or instructions given to the subjects, hence the resulting saliency map cannot accurately reflect the bottom-up, stimulus-driven attention observed under free-viewing conditions.

To have a better understanding of viewers' gaze behaviour when watching videos of varying degrees of perceived quality, a large-scale eye-tracking study was conducted [6]. In this study, a refined experimental methodology was established to ensure the reliable collection of ground truth eye-tracking for video quality perception research. This methodology focuses on capturing saliency, which represents free-viewing, stimulus-driven, bottom-up attention, for both pristine images and their distorted formats of various types and levels of degradation. The experimental conditions and requirements implemented in this study effectively eliminate subject bias caused by stimulus repetition. This approach ensures the collection of a highly reliable eye-tracking data, resulting in the creation of the SVQ160 database. The eye-tracking study involved 160 human observers and 160 video stimuli, providing the best-of-its-kind data for the differences in fixation deployment when viewing pristine/undistorted versus distorted video content. Although the statistical analysis performed in [6] has demonstrated the significance of saliency shifts caused by the distortions appearing in a video, it remains largely unexplored how these distortions alter gaze allocation depending on video content and passage of time. Previous studies have shown that transformations and distortions in static images can affect the distribution of visual attention. For example, in [25], an eye-tracking experiment was conducted using 288 images distorted with five different types of artifacts at three levels of degradation. Similarly, in [19], an eye-tracking dataset was created for over 1,900 images degraded by 19 different types of transformations. By analysing eye movement patterns, both studies reveal that image transformations/distortions cause shifts in the viewers' gaze allocation. However, these studies do not address distortion-induced saliency shifts (DSS) in dynamic video content, which involves both spatial and temporal aspects of saliency influenced by video distortions. Also, a systematic characterisation of DSS as a measurable behavioural response remains unexplored. Developing computational models to simulate DSS behaviours, and consequently integrating DSS into video algorithms would further advance the file.

Additionally, it would be beneficial to build computational models for the behaviours of distortion-induced saliency shifts (DSS), and consequently to integrate DSS into video algorithms.

III. Video Distortion-Induced Saliency Shifts

A. *SVQ160 Database*

The SVQ160 database [6] is so far the largest and most reliable eye-tracking database available in the literature for studying video distortion/quality perception. To create the SVQ160 database, the set of video stimuli was purposely taken from one of the video quality assessment benchmarks i.e., the LIVE database [26], where the ratings of perceived quality are readily available. There are 10 pristine/undistorted reference videos of high quality and their 150 distorted versions of varying perceived quality, i.e., each reference corresponds to 15 distorted videos created by different distortion types including MPEG-2 compression, H.264 compression, simulated transmission of H.264 compressed bitstreams through error-prone IP networks and through error-prone wireless networks. Fig.1 illustrates the 10 reference videos namely 'bs (Blue Sky)', 'mc (Mobile & Calendar)', 'pa (Pedestrian Area)', 'pr (Park Run)', 'rb (River Bed)', 'rh (Rush Hour)', 'sf (Sunflower)', 'sh (Shields)', 'st (Station)' and 'tr (Tractor)' (one frame of each reference video) and the saliency maps rendered from the eye-tracking data [6]. A description of these videos is provided in Table.I. The resolution of all videos is 768 × 432 pixels and the duration is 10 seconds except for the Blue Sky sequence being 8.68 seconds. A rigorously designed and fully controlled eye-tracking experiment was conducted to collect eye movement data for the 160 video stimuli. Each video received unbiased fixations from 20 human observers. The saliency data saturation analysis in [6] demonstrated that gaze patterns reached high consistency with 15 participants and achieved full saturation with 18 participants. The SVQ160 database [6] utilised eye-tracking data from 20 participants, ensuring the generation of saturated and reliable saliency maps.

TABLE I
Description of video stimuli [26].

Video name	Content description
Rush hour (rh)	'Still camera, shows rush hour traffic on a street'
River Bed (rb)	'Still camera, shows a river bed containing some pebbles and water'
Shields (sh)	'Camera pans at first, then becomes still and zooms in; shows a person walking across a display pointing at it'
Blue Sky (bs)	'Circular camera motion showing a blue sky and some trees'
Mobile & Calendar (mc)	'Camera pan, toy train moving horizontally with a calendar moving vertically in the background'
Tractor (tr)	'Camera pan, shows a tractor moving across some fields'
Pedestrian area (pa)	'Still camera, shows some people walking about in a street intersection'
Park run (pr)	'Camera pan, a person running across a park'
Station (st)	'Still camera, shows a railway track, a train and some people walking across the track'
Sunflower (sf)	'Still camera, shows a bee moving over a sunflower in close-up'

B. *Distortion-Induced Saliency Shifts (DSS)*

In this paper, we formulate a quantitative variable of video distortion-induced saliency shifts (DSS). This measure can quantify the difference in visual saliency between the reference/undistorted and distorted videos. To this end, a frame-level saliency map (FSM) is first generated, representing the

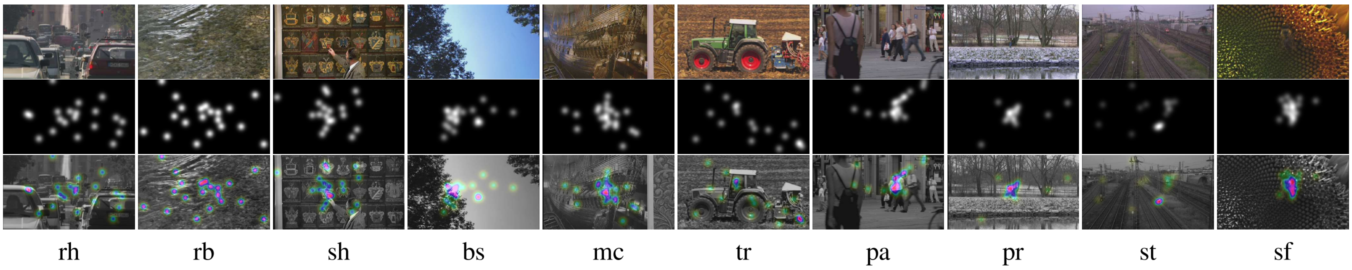


Fig. 1. SVQ160 database: first row illustrates content (representative frames) of the original videos, second row shows saliency maps; and third row shows the corresponding heatmaps (saliency maps superimposed on representative frames).

probability distribution of fixations over one frame within the context of watching a video sequence:

$$\mathrm{FSM}_{(x,y)} = \sum_{i=1}^{N} \exp\left[-\frac{(x_i - x)^2 + (y_i - y)^2}{2\sigma^2}\right], \quad (1)$$

where (x_i, y_i) denotes the pixel position of the i-th fixation, N is the total number of fixations obtained over all subjects in the eye-tracking experiment. The activity of the patch is modelled as a Gaussian distribution, of which the width σ approximates the size of the fovea (i.e., 2° of visual angle, and here σ is equal to 45 pixels in our study). Then, the distortion-induced saliency shifts (DSS) – the difference in fixation distribution between the reference/undistorted frame and distorted frame – are calculated using Pearson Linear Correlation Coefficient (CC). It should be noted that CC has been proven to be the best perception-based metric to assess the saliency difference [16], and is defined as:

$$\mathrm{CC}_{(\mathrm{FSM}\_ref, \mathrm{FSM}\_dis)} = \left|\frac{\mathrm{cov}(\mathrm{FSM}\_ref, \mathrm{FSM}\_dis)}{\sigma_{\mathrm{FSM}\_ref} \times \sigma_{\mathrm{FSM}\_dis}}\right|, \quad (2)$$

where σ_{SM_ref} and σ_{SM_dis} denote the standard deviation of FSM_ref and FSM_dis respectively. Where FSM_ref is the frame-level saliency map of the reference video, and FSM_dis is the frame-level saliency map of the distorted video. $cov(\mathrm{FSM}\_ref, \mathrm{FSM}\_dis)$ represents the covariance. When using CC to measure the similarity between saliency maps, its absolute value is generally considered, ranging from 0 to 1. A CC value closer to 1 indicates a higher similarity between the saliency maps, while a value closer to 0 reflects lower similarity. Finally, DSS can be characterized based on (2) for each video, using the statistics of frame-based CC over time to reveal the spatio-temporal properties of saliency shifts.

IV. Perceptual Behaviour Models of Distortion-Induced Saliency Shifts

Eye movements provide rich information on viewers' cognitive-perceptual behaviours, therefore, incorporating saliency in a video algorithm is of a significant trend in video processing research [7], [17]. The DSS reflects the impact of visual distortions on viewers' attention, contributing to the perception of a distorted video. Building on the statistical hypotheses derived from eye-tracking data, we model human behavioural responses to video distortions using DSS and propose a new behaviour fusion model, named the DSS behaviour fusion (DBF) model. Fig.6 illustrates the proposed DBF model, which comprises three quantitative vision models: video content classifier, time-series weighting function, and DSS-critical frame classifier. The details are described below.

A. Quantitative Vision Models

1) Video content classifier: In [27], a statistical analysis was conducted to verify that video content (VC) has a significant impact on the distortion induced saliency shifts (DSS), where the VC variable is defined as the degree of spatial saliency dispersion (i.e., contracted saliency (VC_compact) and dispersed saliency (VC_dispersed)). The hypothesis testing revealed that "*when watching the distorted videos, the degree of saliency shifts of VC_dispersed is statistically significantly higher than that of VC_compact, relative to the original video content.*" The evidence suggests these two different categories of video content should be separately considered for the VQA metrics that include the saliency component. To this end, a video content classifier is constructed to distinguish between the content with concentrated saliency and content with dispersed saliency, using the multilevel entropy (ME) [28]. For the saliency map (S) of a frame, the frame-level ME is calculated based on Shannon entropy applied to $P \times P$ non-overlapping blocks of S:

$$\mathrm{ME} = H_{\Sigma}(S) = \frac{1}{P^2} \sum_{B=1}^{P^2} H(B), \quad (3)$$

where H represents the entropy of a 2-D image block, P refers to the segmentation level (i.e., $P = 4$ is empirically determined in [28] and used here), and B runs over each block.

Lower ME values indicate saliency is concentrated in fewer regions, while higher entropy suggests it is dispersed throughout the spatial domain. For each video, the sequence-level ME (SME) is calculated by taking the average of frame-level ME values. For a given video (v), the video content classifier can be expressed as:

$$Cv = \begin{cases} v \in \ VC\_dispersed & \text{if } SME(v) \geq \tau_{SME}(vt), \\ v \in \ VC\_compact & \text{if } SME(v) < \tau_{SME}(vt), \end{cases} \quad (4)$$

where $\tau_{SME}(vt)$ denotes the threshold for the classifier. In this study, the threshold is determined by ranking the original videos as per the SME, and from the highest to lowest finding the first pair of adjacent videos that have a statistically significant difference in their SME values. Fig.2 illustrates the SME value (as well as the saliency map of a representative frame) for each original video contained in the SVQ160 database.

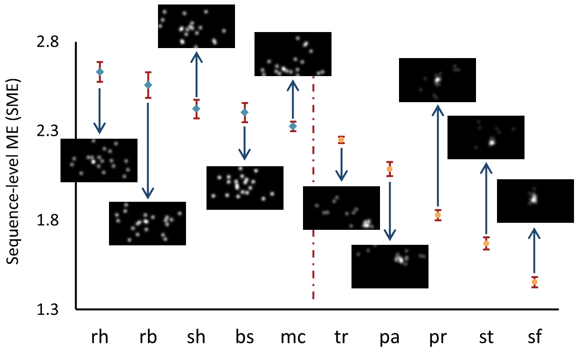
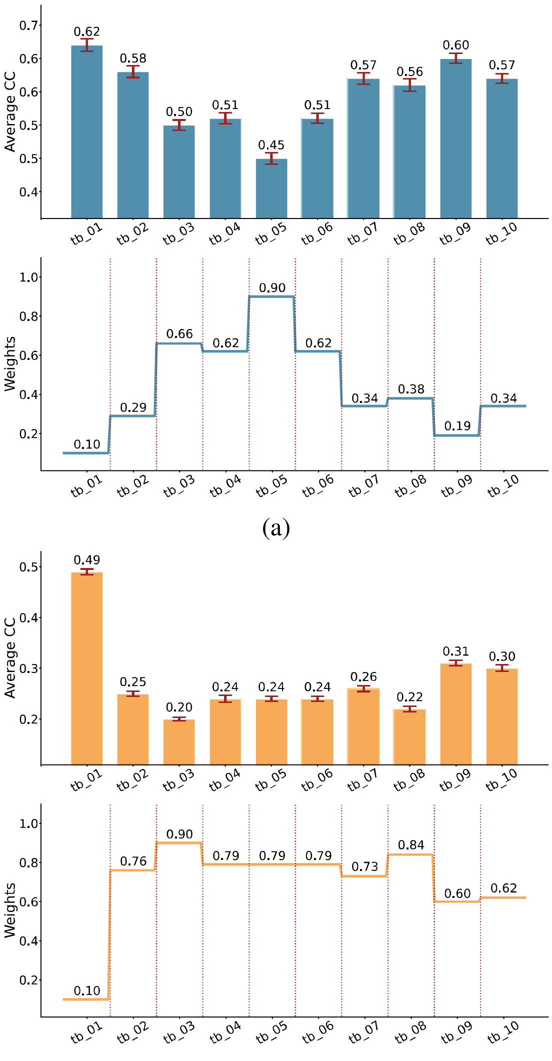


Fig. 2. Illustration of the SME value (as well as the saliency map of a representative frame) for each original video contained in the SVQ160 database. The error bars indicate a 95% confidence interval.

2) Time-series weighting function: In [27], a statistical analysis was conducted to verify that time order (TO) has a significant impact on the DSS, where the duration of a video is divided into successive blocks of time. By formulating the TO into three semantic categories including TO_beginning, TO_middle, and TO_end, the hypothesis testing revealed that "*there is a significant difference between TO_beginning, TO_middle and TO_end; viewers' gaze is less affected by distortions in the beginning of video playback than that in the rest of viewing time; in the middle of viewing there are significant saliency shifts due to the occurrence of distortions; the impact of distortions on gaze behaviour significantly decreases towards the end of viewing.*" The evidence reflects the viewers' sensitivity to distortions in different viewing periods, which may be affected by e.g., the centre-bias effect [29] in the beginning of the scene, and learning to tolerate the distortions from middle to the end of viewing. Based on this, a time-series weighting function is constructed by fitting a step function to approximate the gaze behaviour. We divide the time duration into 10 successive blocks of time each representing one second of video playback. Within each time block, we calculate the average of the frame-level CC values (i.e., the measure of DSS, see (2)) over two sets of videos (i.e., classified as VC_compact and VC_dispersed) contained in the SVQ160 database, respectively. The results enable us to fit two step functions, one for videos of concentrated saliency and one for videos of dispersed saliency, approximating the temporal progression of DSS behaviours as shown in Fig.3. In constructing the weighting function, we consider the impact of DSS on video quality assessment (VQA) - the time blocks during which viewers are more sensitive to distortions hold greater significance in VQA. Hence, we assign larger weights to the more DSS affected frames and smaller weights to less DSS affected frames of a video in the calculation of a VQA metric. For each time block (tb), the mapping from the average of the frame-level CC values to the weight is expsressed as:

$$W_t(i) = F_{\text{scale}}\left(1 - CC_{tb}(i)\right), \tag{5}$$



(a)

(b)

Fig. 3. Temporal progression of DSS behaviours and time-series weighting function for (a) videos of concentrated saliency and (b) videos of dispersed saliency. The error bars indicate a 95% confidence interval.

where $W_t(i)$ represents the weight for the i-th time block; $CC_{tb}(i)$ is the average of the frame-level CC values for the i-th time block. F_{scale} normalises the value to be within the range between 0.1 and 0.9, eliminating the occurrence of 0 and 1 in the weighting function for practical purposes:

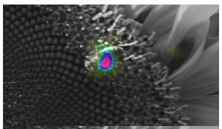
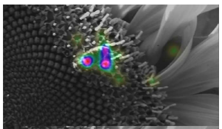
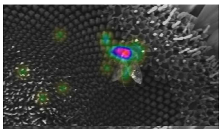
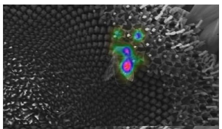
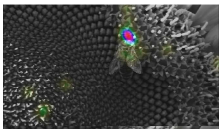
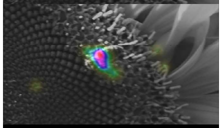
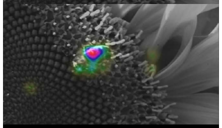
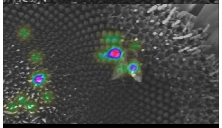
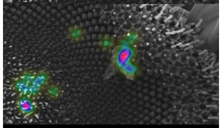
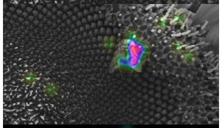
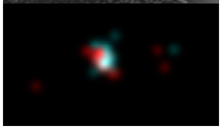
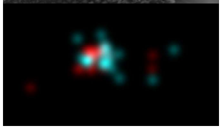
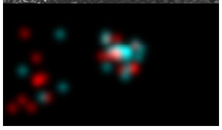
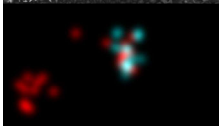
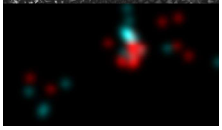
$$F_{\text{scale}}(x) = 0.1 + \frac{F(x) - \min(x)}{\max(x) - \min(x)} \times 0.8. \tag{6}$$

This results in the weighting function W_{tc} for content of concentrated saliency and W_{td} for content of dispersed saliency.

3) DSS-critical frame classifier: The way humans perceive and judge the overall quality of a video highly depends on their behavioural responses to the time-varying distortion profile [27]. The frames in a distorted video that exhibit significant gaze shifts relative to the pristine video are perceptually critical [17] and largely determine the summation strategy of distortions in the spatio-temporal domain. We refer to these frames as DSS-critical frames and develop a method to extract these frames from a distorted video. The measure CC (see (2)) has been proven the most effective measure

TABLE II

ILLUSTRATIONS OF DISTORTION INDUCED SALIENCY SHIFTS (DSS) AND COMPARISON OF SENSITIVITY OF MEASUREMENT USING PEARSON LINEAR CORRELATION COEFFICIENT (CC) AND FRÉCHET DISTANCE (FD). SOME REPRESENTATIVE FRAMES SPANNING THE ENTIRE VIDEO DURATION ARE EXTRACTED FROM A PRISTINE VIDEO (I.E., SUNFLOWER) AND ITS DISTORTED VERSION, AND THE CORRESPONDING GROUND-TRUTH SALIENCY MAPS ARE SUPERIMPOSED ON THESE FRAMES. ALSO, THE SALIENCY MAP OF THE PRISTINE FRAME AND THE SALIENCY MAP OF THE DISTORTED FRAME ARE SUPERIMPOSED TO ILLUSTRATE THE DIFFERENCE IN SALIENCY PATTERN ALLOCATION.

Reference frames					
Distorted frame (MPEG) DMOS: 40.999					
Superimposition of saliency maps					
Frame number	20	29	130	146	229
CC value	0.878	0.692	0.630	0.554	0.460
FD value	9560.567	9765.978	3323.546	55200.28	6991.081

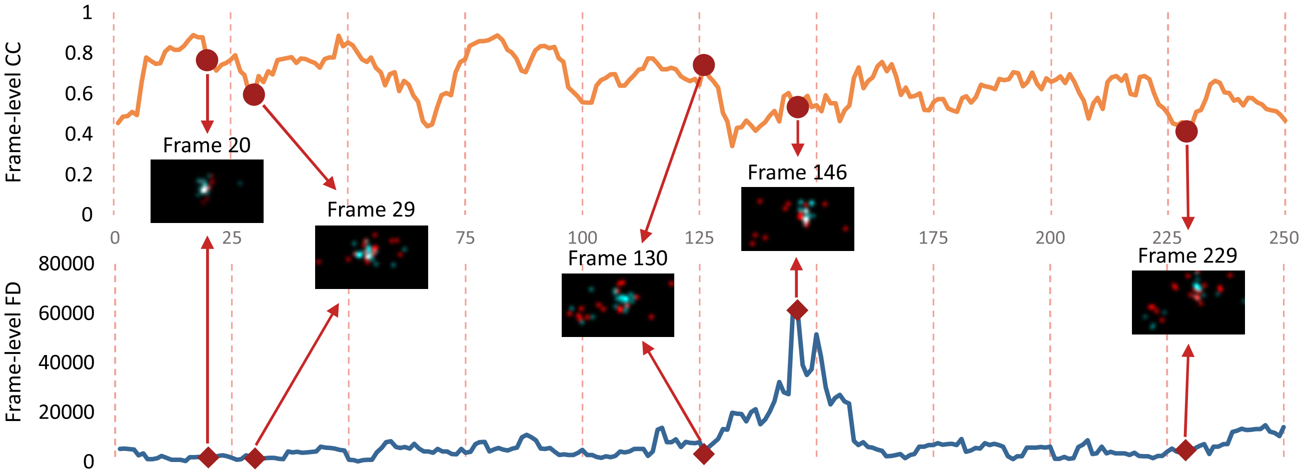


Fig. 4. Illustrates of the comparison of temporal profiles using the frame-level CC and FD for the same videos (reference and distorted) in Table II.

for saliency difference, however, it primarily captures the perceived difference of global saliency patterns and does not adequately respond to the difference of the localised patterns. Therefore a more sensitive measure is required to effectively identify frames that cause the pronounced saliency shifts in a distorted video. In light of this, we adopt the Fréchet distance (FD), which measures the difference between two probability distributions. We consider the saliency patterns of the original frame and that of the distorted frame as two distributions, FD is implemented as follows:

$$FD = \|\mu_r - \mu_d\|_2^2 + \operatorname{tr}\left(\sigma_r + \sigma_d - 2\left(\sigma_r \sigma_d\right)^{\frac{1}{2}}\right), \quad (7)$$

where μ_r and μ_d denote the average feature distribution of the saliency maps; σ_r and σ_d denote covariance matrices. The greater value of FD means that the gaze shifts are more substantial.

As shown in Table II, some representative frames spanning the entire video duration are extracted from a pristine video and its distorted version in our study, and the corresponding ground-truth saliency maps are superimposed on these frames. Also, the saliency map of the pristine frame and the saliency map of the distorted frame are superimposed to illustrate the difference in saliency pattern allocation. The difference is measured by CC and FD, respectively. It clearly demonstrates the superiority of FD over CC in distinguishing between the pronounced saliency shifts and less pronounced saliency shifts. For example, the CC values (0.69 versus 0.63) for frame 29 and frame 130 are similar albeit the degrees of DSS are observed rather different in the superimposition visualisations of saliency maps; the disparity in FD values (9765 versus 3323) better reflects the difference. It is evident that compared to CC measure, FD measure provides a greater level of sensitivity to changes of saliency patterns.

Fig.4 further illustrates the comparison of temporal profiles using the frame-level CC and FD, respectively, for the same videos (reference and distorted) in Table II. It can be seen from the figure that the CC-based profile is not sensitive

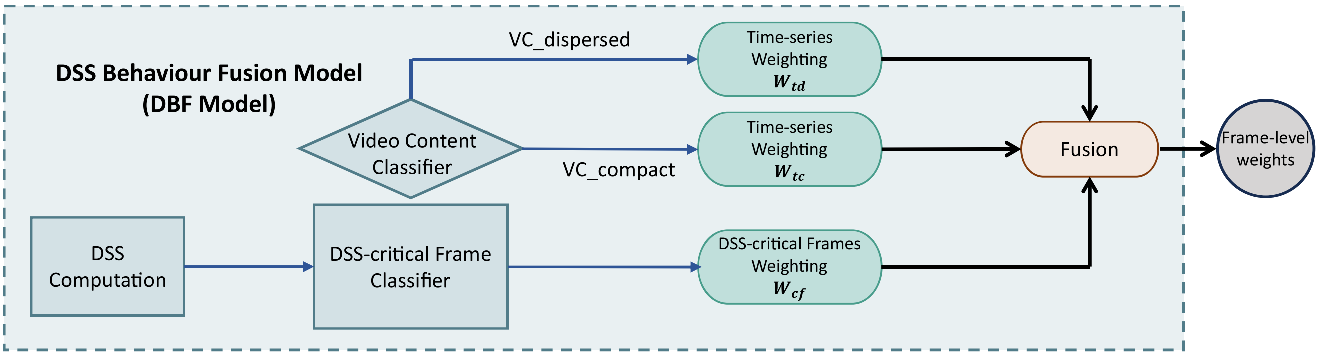

DSS Behaviour Fusion Model
(DBF Model)
VC_dispersed
Time-series Weighting W_{td}
Video Content Classifier
VC_compact
Time-series Weighting W_{tc}
Fusion
Frame-level weights
DSS Computation
DSS-critical Frame Classifier
DSS-critical Frames Weighting W_{cf}

Fig. 5. Illustration of the proposed DSS behaviour fusion (DBF) model. It consists of the computation of the distortion-induced saliency shifts (DSS) and the ensemble of quantitative vision models to generate the frame-level weights for a distorted video.

to the highly DSS affected frames (i.e., a small disparity between the pronounced saliency shifts and less pronounced saliency shifts), while the FD-based profile can well capture the pronounced saliency shifts (i.e., a peak occurs in the plot). This suggests that FD can be used as a metric to identify the DSS-critical frames, which have a disruptive impact on assessing the overall video quality. We construct a simple classifier to identify DSS-critical frames in a video j, using a threshold:

$$\tau_{fd}(j) = \mu_{fd}(j) + 2 \times \sigma_{fd}(j), \tag{8}$$

where $\mu_{fd}(j)$ and $\sigma_{fd}(j)$ denote the mean and standard deviation of the video's frame-level FD values.

Each video frame is classified by the threshold τ_{fd}. Based on this classification, a set of frame-level weights Wcf is then derived to express the level of DSS disruption of each frame on a VQA metric:

$$W_{cf} = \begin{cases} \alpha & \text{if } FD(k) \geq \tau_{fd}(j), \\ 1-\alpha & \text{if } FD(k) < \tau_{fd}(j), \end{cases} \tag{9}$$

where α is a DSS disruption weighting factor, k denotes the k-th frame of video j. In this paper, based on the empirical experiments, α is set to be 0.6.

B. DSS Behaviour Fusion Model

Now we assemble the quantitative vision models to form a DSS behaviour fusion (DBF) model as illustrated in Fig.5, taking into account the spatial and temporal impacts of DSS on the perception of a distorted video. The DBF model consists of two parallel pipelines: time-series weighting and DSS-critical frame weighting. For the time-series weighting, the input video, based on its saliency, is firstly classified into content of concentrated saliency and content of dispersed saliency, and then a specific time-series weighing function (i.e., W_{tc} or W_{td}) is determined accordingly. For the DSS-critical frame weighting, the input video is first analysed based on its saliency frame-by-frame to identify DSS-critical frames in the video, and then a weighting function W_{cf} is applied. Finally, a sequence of frame-level weights W_f is generated via a fusion operation of the results of time-series weighting and DSS-critical frame weighting as follows:

$$W_f = \begin{cases} W_{td} \times W_{cf} & \text{if } VC\_dispersed, \\ W_{tc} \times W_{cf} & \text{if } VC\_compact. \end{cases} \tag{10}$$

The weighting effectively combines the results of the three vision models for DSS, thereby integrating the spatial and temporal impacts of DSS on video distortion perception. The resulting frame-level weights for a video comprehensively reflect viewer's behavioural responses to distortions at different viewing times as well as to disruptive distortions throughout the entire viewing experience.

V. Application of Distortion-Induced Saliency Shifts

Distortion-induced saliency shifts (DSS) measure how alterations in a video's properties can change the perceived importance of different regions in the spatial and temporal domains. By leveraging this concept, video algorithms can be perceptually optimised to align with viewers' attention. The computational model, i.e., the DSS behaviour fusion (DBF) model, has promising applications in various video algorithms, enhancing both efficiency and effectiveness. For example, in video compression, understanding how distortions influence saliency allows algorithms to prioritize and preserve regions that are visually attended by viewers, achieving higher compression rates without compromising perceptual quality. In video enhancement, adaptive noise reduction can be utilised to optimise the algorithm, improving the effectiveness of content enhancement. In this paper, we focus on the application of DSS in the emerging field of video quality assessment (VQA).

A. Validation of Effectiveness of DSS

Video quality assessment (VQA) often takes advantage of the well-studied image quality measures to quantify the perceived quality of individual frames in a video, and produce a sequence-level quality measure using sophisticated temporal pooling [30]. However, the main challenges remain as to how humans respond perceptually to the visual distortions when watching a video, and how to quantify these behavioural responses in a VQA algorithm. To validate the effectiveness of the proposed DSS behaviour fusion (DBF) model, we establish a DSS-based VQA framework, as shown in Fig.6. The computational framework incorporates our proposed DSS

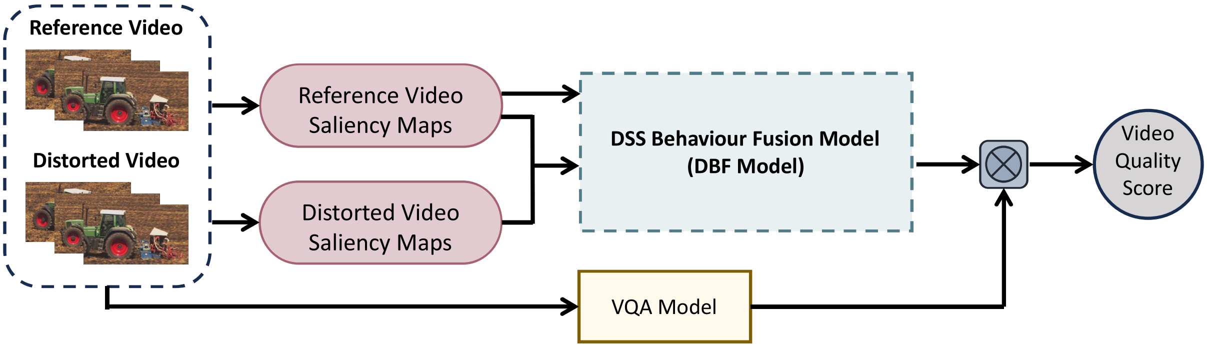

Reference Video
Distorted Video
Reference Video Saliency Maps
Distorted Video Saliency Maps
DSS Behaviour Fusion Model (DBF Model)
VQA Model
Video Quality Score

Fig. 6. Illustration of the proposed DSS-based video quality assessment (VQA) framework. It integrates the proposed DSS behaviour fusion (DBF) model as a self-contained unit into a VQA model. This integration allows quantifying the contribution of DSS through the performance gain achieved by the VQA model.

TABLE III
Performance gain of state-of-the-art VQA models by integrating the proposed DSS behaviour fusion (DBF) model. In statistical significance, "1" means that the performance gain is statistically significant; "0" means that the performance gain is not significant. The saliency maps used in the DBF model are the ground truth contained in the SVQ160 database.

VQA	Original Performance		Performance Gain (+DBF Model))		Sig.
	PLCC	SROCC	ΔPLCC	ΔSROCC	
PSNR [31]	0.4718	0.4590	**4.39%↑**	**6.97%↑**	1
SSIM [32]	0.5115	0.5164	**2.42%↑**	**1.66%↑**	1
MS-SSIM [33]	0.6950	0.7331	**1.10%↑**	**1.27%↑**	1
VIF [34]	0.6402	0.6491	**2.45%↑**	**1.38%↑**	1
VIFP [34]	0.6402	0.6491	**2.45%↑**	**1.38%↑**	1
GMSD [35]	0.7461	0.7362	**2.26%↑**	**2.33%↑**	1
GMSM [35]	0.6142	0.6676	**2.81%↑**	**1.47%↑**	1
SpEED [36]	0.5826	0.6203	**1.98%↑**	**1.12%↑**	1
VMAF [37]	0.7347	0.7555	**0.80%↑**	**0.57%↑**	1

TABLE IV
Performance gain of state-of-the-art VQA models by integrating the proposed DSS behaviour fusion (DBF) model. In statistical significance, "1" means that the performance gain is statistically significant; "0" means that the performance gain is not significant. The saliency maps used in the DBF model are automatically generated by a computational model (i.e., ViNet [38]).

VQA	Original Performance		Performance Gain (+DBF Model))		Sig.
	PLCC	SROCC	ΔPLCC	ΔSROCC	
PSNR [31]	0.4718	0.4590	**2.96%↑**	**0.75%↑**	1
SSIM [32]	0.5115	0.5164	**1.10%↑**	**2.53%↑**	1
MS-SSIM [33]	0.6950	0.7331	**1.04%↑**	**1.18%↑**	1
VIF [34]	0.6402	0.6491	**1.69%↑**	**1.53%↑**	1
VIFP [34]	0.6402	0.6491	**1.69%↑**	**1.53%↑**	1
GMSD [35]	0.7461	0.7362	**0.26%↑**	**0.31%↑**	1
GMSM [35]	0.6142	0.6676	**1.49%↑**	**1.35%↑**	1
SpEED [36]	0.5826	0.6203	**1.07%↑**	**1.36%↑**	1
VMAF [37]	0.7347	0.7555	**0.67%↑**	**1.10%↑**	1

behaviour fusion (DBF) model as a self-contained unit into an existing VQA model. This integration allows us to quantify the contribution of DSS through the performance gain achieved by the VQA model. It should be noted that as per the nature of DSS, the VQA models selected in this study must meet the following criteria: (1) a full-reference VQA model which uses the reference/pristine video and the distorted video to measure the perceived quality of the distorted video; and (2) a frame-level VQA model which explicitly calculates quality scores of individual frames in the distorted video sequence. Also, to have a comprehensive evaluation, we include both traditional VQA models that are based on pixel-based or hand-crafted visual features; and learning-based VQA models that adopt machine learning techniques or deep neural networks to learn visual representations. Finally, we selected nine widely used VQA methods including PSNR [31], SSIM [32], MS-SSIM [33], VIF [34], VIFP [34], GMSD [35], GMSM [35], SpEED [36] and VMAF [37]. The quality prediction performance of a VQA model is quantified by the Pearson's correlation coefficient (PLCC) and Spearman's rank correlation coefficient (SROCC).

For each VQA model, we produce its DSS-based version (i.e., by integrating the DBF model) using the framework as shown in Fig.6, and we compare the quality prediction performance of the original VQA model versus the DSS-based VQA model on the SVQ160 database. Since the SVQ160 database is the only database available in the literature that contains both the ground truth saliency and subjective quality scores, the evaluation using this database can faithfully reveal the added value of the proposed DBF model. Table III shows the performance gains of VQA models by adding the DSS component. It can be seen that there is a gain in quality prediction performance when using the proposed DBF model. To verify whether the performance gain is statistically significant, hypothesis testing in conducted. As prescribed in [39], the statistical test is based on the residuals between the subjective quality scores (i.e., MOS) and the predictions of a VQA (i.e., either the original VQA model or DSS-based VQA model). We first evaluate the assumption of normality of the residuals ($\|MOS - VQA\|$ and $\|MOS - DSS\_VQA\|$), when paired residuals are both normal, a paired samples t-test is performed; otherwise a non-parametric test i.e., Wilcoxon signed rank test is performed. The significance test results are shown in Table III, which verifies that in all cases the gain in performance is statistically significant. This demonstrates the added value of using the proposed DBF model for video quality assessment, and the importance of incorporating viewers' gaze behaviour in perceptual tasks.

Realistically, ground truth saliency is often not available, and a practical solution is to calculate saliency by a computational model. To investigate whether the model-generated

TABLE V
ABLATION STUDY: A VARIANT OF THE DSS-BASED VQA FRAMEWORK CONSTRUCTED SPECIFICALLY TO VERIFY THE CONTRIBUTION OF KEY COMPONENTS OF THE PROPOSED DSS BEHAVIOUR FUSION (DBF) MODEL. (**BOLD** FONT MEANS THAT THE DSS-BASED VQA'S PERFORMANCE GAIN IS THE LARGEST IN THE ABLATION STUDY.)

W_t	W_{cf}	PSNR [31]	SSIM [32]	MS-SSIM [33]	VIF [34]	VIFP [34]	GMSD [35]	GMSM [35]	SpEED [36]	VMAF [37]
–	–	0.4718	0.5115	0.6950	0.6402	0.6402	0.7461	0.6142	0.5826	0.7347
✓	–	2.15%↑	1.22%↑	0.89%↑	0.81%↑	0.81%↑	1.06%↑	1.31%↑	0.76%↑	0.48%↑
–	✓	0.52%↑	0.12%↑	0.10%↑	0.14%↑	0.14%↑	0.08%↑	0.09%↑	0.07%↑	0.12%↑
✓	✓	**4.39%↑**	**2.42%↑**	**1.10%↑**	**2.45%↑**	**2.45%↑**	**2.26%↑**	**2.81%↑**	**1.98%↑**	**0.80%↑**

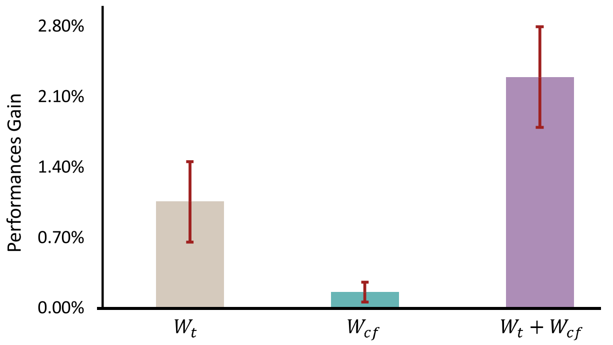


Fig. 7. Ablation study: Performance gain (in terms of PLCC) averaged over all VQA models using three variants of the DSS-based VQA framework, i.e., W_t, W_{cf} and $W_t + W_{cf}$. The error bars indicate a 95% confidence interval.

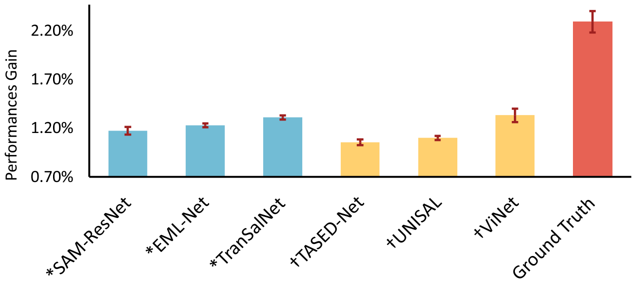


Fig. 8. Ablation study: Performance gain (in terms of PLCC) averaged over all VQA models for different saliency models (including the ground truth) used to produce saliency maps for the DBF model. The error bars indicate a 95% confidence interval. * and † represent static and dynamic saliency prediction models, respectively.

saliency when used as a substitute for ground truth saliency can still bring in VQA performance gain, we repeat the above experiment protocol using a state-of-the-art video saliency prediction model (i.e., ViNet [38]) instead of the ground truth saliency of the SVQ160 database. The results shown in Table IV demonstrate that the effectiveness of a computational saliency model for the proposed DSS-based framework; and that the performance gain is statistically significant. The machine-generated saliency makes the DSS-based VQA more applicable for real-world scenarios.

B. Ablation Study

Ablation studies are conducted to verify the contribution of key components of the proposed DSS behaviour fusion (DBF) model, including (1) the time-series weighting (i.e., W_t), and (2) the DSS-critical frame weighting (i.e., W_{cf}). Table V shows the performance of VQA models without using the DSS and with three variants of using the DSS weightings (i.e., W_t, W_{cf} or $W_t + W_{cf}$). It is evident that the application of W_t and W_{cf} individually contributes to performance enhancement for the nine VQA models; and that the appellation of W_t and W_{cf} combined together provides the largest performance gain for these VQA models, as shown in Fig.7. This suggests the efficacy of the proposed DSS behaviour fusion (DBF) model.

In addition, we evaluate the relative impact of different saliency prediction models on the computational DSS-based VQA framework, which integrates our proposed DBF model in to a VQA. To this end, we select six state-of-the-art saliency models including EML-NET [41], TranSalNet [42], SAM-ResNet [40], ViNet [38], UNSIAL [44], and TASED-Net [43]; and implement them to generate saliency maps of videos contained in the SVQ160 database. The first three models are static models designed for predicting visual saliency of images, and the last three are dynamic models specifically designed for predicting saliency of videos. These models have been proven the best-performing models as per the widespread saliency benchmarks [18]. Tables VI illustrates the performance of nine DSS-based VQA models based on different saliency prediction models. In general, these state-of-the-art saliency models can be effectively embedded in the proposed DSS-based VQA framework to enhance the VQA models' performance. Fig.8 shows the performance gain averaged over all VQA models for different saliency models including the ground truth. It can be seen that TranSalNet and ViNet provide the largest gain compared to other saliency models. But there is still room for improvement compared to the gain produced by the ground truth saliency.

VI. DISCUSSION

In this study, model development and validation were performed using the SVQ160 dataset [6], which currently serves as the only dataset combining both eye movements and subjective video quality assessments. Evaluating the proposed approach on an independent dataset is essential for assessing its generalizability and robustness. To further enhance the model's applicability, future work should involve conducting additional experiments to collect eye movements and subjective video quality assessments from more diverse video datasets, incorporating a wider range of video resolutions, content types, and distortion scenarios. Such efforts will help ensure that the proposed DSS models generalise effectively across different datasets and real-world application contexts.

While our current study primarily focuses on measuring distortion-induced saliency shifts (DSS) and applying DSS to video algorithm optimisation, we acknowledge the potential impact of high-resolution video on DSS. As resolution can influence viewers' perception in general, higher resolutions

TABLE VI

Ablation study: Performance gain of VQA models by integrating the proposed DSS behaviour fusion (DBF) model. The saliency maps used in the DBF model can be automatically generated by different saliency prediction models including SAM-ResNet [40], EML-NET [41], TranSalNet [42], TASED-Net [43], UNSIAL [44], and ViNet [38]. * and † represent static and dynamic saliency prediction models, respectively.

Models		PSNR	SSIM	MS-SSIM	VIF	VIFP	GMSD	GMSM	SpEED	VMAF
		[31]	[32]	[33]	[34]	[34]	[35]	[35]	[36]	[37]
Original	PLCC	0.4718	0.5115	0.6950	0.6402	0.6402	0.7461	0.6142	0.5826	0.7347
	SROCC	0.4590	0.5164	0.7331	0.6491	0.6491	0.7362	0.6676	0.6203	0.7555
*SAM-ResNet	ΔPLCC	**3.53%**↑	**2.22%**↑	**0.03%**↑	**1.19%**↑	**1.19%**↑	**1.39%**↑	**1.04%**↑	-0.40%↓	**0.38%**↑
(DBF)	ΔSROCC	**0.01%**↑	**2.49%**↑	**0.86%**↑	**0.95%**↑	**0.95%**↑	**2.03%**↑	**2.07%**↑	**0.46%**↑	**0.82%**↑
*EML-NET	ΔPLCC	**1.16%**↑	**1.31%**↑	**0.74%**↑	**1.22%**↑	**1.22%**↑	**0.83%**↑	**2.24%**↑	**1.64%**↑	**0.72%**↑
(DBF)	ΔSROCC	**1.83%**↑	**0.73%**↑	**0.85%**↑	**1.45%**↑	**1.45%**↑	**0.88%**↑	**0.83%**↑	**0.87%**↑	**0.38%**↑
*TranSalNet	ΔPLCC	**4.15%**↑	**1.39%**↑	**0.79%**↑	**0.40%**↑	**0.40%**↑	**0.93%**↑	**1.81%**↑	**1.55%**↑	**0.37%**↑
(DBF)	ΔSROCC	**6.49%**↑	**2.18%**↑	**1.47%**↑	**0.92%**↑	**0.92%**↑	**1.02%**↑	**0.98%**↑	-0.39%↓	**0.32%**↑
†TASED-Net	ΔPLCC	**3.63%**↑	**0.12%**↑	**0.54%**↑	-0.11%↓	-0.11%↓	**1.09%**↑	**2.74%**↑	**2.14%**↑	-0.54%↓
(DBF)	ΔSROCC	**4.51%**↑	**0.81%**↑	**0.24%**↑	**0.04%**↑	**0.04%**↑	**0.07%**↑	**0.40%**↑	**0.52%**↑	**0.43%**↑
†UNISAL	ΔPLCC	**2.69%**↑	**1.27%**↑	-0.21%↓	-0.09%↓	-0.09%↓	**0.69%**↑	**2.47%**↑	**2.69%**↑	**0.47%**↑
(DBF)	ΔSROCC	**0.88%**↑	**2.82%**↑	**0.08%**↑	**0.84%**↑	**0.84%**↑	**0.49%**↑	-0.41%↓	**0.72%**↑	**0.05%**↑
†ViNet	ΔPLCC	**2.96%**↑	**1.10%**↑	**1.04%**↑	**1.69%**↑	**1.69%**↑	**0.26%**↑	**1.49%**↑	**1.07%**↑	**0.67%**↑
(DBF)	ΔSROCC	**0.75%**↑	**2.53%**↑	**1.18%**↑	**1.53%**↑	**1.53%**↑	**0.31%**↑	**1.35%**↑	**1.36%**↑	**1.10%**↑

may reveal finer visual details, making certain distortions more perceptually noticeable and potentially leading to more pronounced spatial DSS. Extending the investigation to high-resolution content to systematically study the impact of resolution on DSS represents an important next step. Future work will involve creating new eye-tracking datasets and refining our computational models to account for the increased complexity of high-resolution visual stimuli. We anticipate that our developed DSS models can be adapted to handle high-resolution content by incorporating appropriate scaling mechanisms and advanced feature extraction techniques, enabling broader applicability of DSS models across diverse video formats.

In visual quality assessment, no-reference algorithms play a critical role by complementing full-reference methods. The no-reference algorithms, such as BRISQUE [45], NIQE [46], BPRI [47], BMPRI [48], and RichIQA [49], estimate perceived quality without requiring a reference, making them essential for real-world applications where reference content is unavailable. These no-reference algorithms primarily target static images but have inspired video quality assessment (VQA) models that operate without reference content. While our current work focuses on full-reference VQA using distortion-induced saliency shifts (DSS), future research could explore adapting DSS-based models for no-reference VQA tasks. This would involve developing ways to model DSS behaviours using only distorted videos, and a more universal version of the DBF framework, capable of enhancing the performance of no-reference VQA algorithms by leveraging implicit distortion-saliency relationships.

In Section V.B, we evaluated six state-of-the-art saliency prediction models within our proposed DBF framework to assess their effectiveness. As illustrated in Figure 8, the static saliency prediction model TranSalNet and the dynamic model Vi-Net showed the highest performance gains, with average improvements of 1.31% and 1.33% across nine VQA algorithms, respectively. Based on their results on SALICON [18] and DHF1K [50] benchmarks, TranSalNet and Vi-Net rank as top-performing models among the three static and three dynamic saliency prediction models evaluated. However, the performance gains achieved with these computational saliency models remain considerably lower than those obtained using ground truth saliency. This implies that the effectiveness of the DBF framework depends heavily on the accuracy of saliency predictions – models that produce saliency maps closer to the ground truth yield better DBF framework performance.

While our current study focuses solely on visual scene and distortions in video, extending the proposed concept of DSS and its models to consider audio information is an important direction for future research. As a critical component of multimedia content, audio cues can influence viewers' visual attention and overall quality experience. Future studies would involve conducting eye-tracking experiments with synchronised audio-visual stimuli and developing multimodal models capable of capturing cross-modal saliency shifts.

VII. Conclusion

In this paper, we introduce a novel concept - distortion-induced saliency shifts (DSS) in video. Based on a large-scale, reliable eye-tracking database, we conduct an exhaustive statistical analysis to conceptualise DSS and its measurement in the context of video distortions. The perceptual behaviours of DSS are modeled to reflect the impact of video content, passage of time and critical distortion disruption on the perception of video distortions. These vision models are integrated to construct a new DSS behaviour fusion (DBF) model for quantifying viewers' spatio-temporal behaviours. By applying the proposed DBF model in the emerging field of video quality assessment (VQA), we demonstrate the added value of DSS in enhancing video algorithms. Future work will focus on the application of DSS in other important video algorithms.

REFERENCES

[1] V. Cisco, "Cisco visual networking index: Forecast and trends, 2017–2022 white paper," *Cisco Internet Rep*, vol. 17, p. 13, 2019.

[2] X. Min, H. Duan, W. Sun, Y. Zhu, and G. Zhai, "Perceptual video quality assessment: A survey," *Sci. China Inf. Sci.*, vol. 67, no. 11, p. 211301, 2024.

[3] S. Wen, L. Yang, M. Xu, M. Qiao, T. Xu, and L. Bai, "Saliency prediction on mobile videos: A fixation mapping-based dataset and a transformer approach," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 7, pp. 5935–5950, 2024.

[4] L. Jiang, M. Xu, Z. Wang, and L. Sigal, "Deepvs2. 0: A saliency-structured deep learning method for predicting dynamic visual attention," *International Journal of Computer Vision*, vol. 129, no. 1, pp. 203–224, 2021.

[5] M. Qiao, Y. Liu, M. Xu, X. Deng, B. Li, W. Hu, and A. Borji, "Joint learning of audio–visual saliency prediction and sound source localization on multi-face videos," *International Journal of Computer Vision*, vol. 132, no. 6, pp. 2003–2025, 2024.

[6] W. Zhang and H. Liu, "Study of saliency in objective video quality assessment," *IEEE Trans. Image Process.*, vol. 26, no. 3, pp. 1275–1288, 2017.

[7] W. Y. Akamine and M. C. Farias, "Video quality assessment using visual attention computational models," *J. Electron. Imaging*, vol. 23, no. 6, p. 061107, 2014.

[8] V. Lyudvichenko, M. Erofeev, A. Ploshkin, and D. Vatolin, "Improving video compression with deep visual-attention models," in *Proc. 2019 Int. Conf. Intell. Med. Image Process.*, 2019, pp. 88–94.

[9] Z. Li, S. Qin, and L. Itti, "Visual attention guided bit allocation in video compression," *Image Vis. Comput.*, vol. 29, no. 1, pp. 1–14, 2011.

[10] G. Zhai and X. Min, "Perceptual image quality assessment: a survey," *Sci. China Inf. Sci.*, vol. 63, pp. 1–52, 2020.

[11] X. Min, K. Gu, G. Zhai, X. Yang, W. Zhang, P. Le Callet, and C. W. Chen, "Screen content quality assessment: Overview, benchmark, and beyond," *ACM Comput. Surv.*, vol. 54, no. 9, pp. 1–36, 2021.

[12] L. Yang, M. Xu, S. Li, Y. Guo, and Z. Wang, "Blind vqa on 360° video via progressively learning from pixels, frames, and video," *IEEE Transactions on Image Processing*, vol. 32, pp. 128–143, 2023.

[13] D. Varga, "No-reference video quality assessment using multi-pooled, saliency weighted deep features and decision fusion," *Sensors*, vol. 22, no. 6, p. 2209, 2022.

[14] J. You, T. Ebrahimi, and A. Perkis, "Attention driven foveated video quality assessment," *IEEE Trans. Image Process.*, vol. 23, no. 1, pp. 200–213, 2013.

[15] X. Guan, F. Li, Y. Zhang, and P. C. Cosman, "End-to-end blind video quality assessment based on visual and memory attention modeling," *IEEE Trans. Multimed.*, 2022.

[16] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "What do different evaluation metrics tell us about saliency models?" *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 3, pp. 740–757, 2018.

[17] A. Borji, D. N. Sihite, and L. Itti, "Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study," *IEEE Trans. Image Process.*, vol. 22, no. 1, pp. 55–69, 2012.

[18] M. Jiang, S. Huang, J. Duan, and Q. Zhao, "Salicon: Saliency in context," in *2015 IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 1072–1080.

[19] W. Zhang and H. Liu, "Toward a reliable collection of eye-tracking data for image quality research: Challenges, solutions, and applications," *IEEE Trans. Image Process.*, vol. 26, no. 5, pp. 2424–2437, 2017.

[20] U. Engelke, H. Liu, J. Wang, P. Le Callet, I. Heynderickx, H.-J. Zepernick, and A. Maeder, "Comparative study of fixation density maps," *IEEE Transactions on Image Processing*, vol. 22, no. 3, pp. 1121–1133, 2013.

[21] C. T. Vu, E. C. Larson, and D. M. Chandler, "Visual fixation patterns when judging image quality: Effects of distortion type, amount, and subject experience," in *2008 IEEE SW Symp. Image Anal. Interpret.* IEEE, 2008, pp. 73–76.

[22] X. Min, G. Zhai, Z. Gao, and C. Hu, "Influence of compression artifacts on visual attention," in *2014 IEEE Int. Conf. Multimed. Expo (ICME)*. IEEE, 2014, pp. 1–6.

[23] J. Redi, H. Liu, R. Zunino, and I. Heynderickx, "Interactions of visual attention and quality perception," in *Hum. Vis. Electron. Imaging XVI*, vol. 7865. SPIE, 2011, pp. 267–277.

[24] O. Le Meur, A. Ninassi, P. Le Callet, and D. Barba, "Overt visual attention for free-viewing and quality assessment tasks: Impact of the regions of interest on a video quality metric," *Signal Process. Image Commun.*, vol. 25, no. 7, pp. 547–558, 2010.

[25] Z. Che, A. Borji, G. Zhai, X. Min, G. Guo, and P. Le Callet, "How is gaze influenced by image transformations? dataset and model," *IEEE Trans. Image Process.*, vol. 29, pp. 2287–2300, 2020.

[26] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "Study of subjective and objective quality assessment of video," *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1427–1441, 2010.

[27] X. Wu, Z. Dong, F. Zhang, P. L. Rosin, and H. Liu, "Analysis of video quality induced spatio-temporal saliency shifts," in *2022 IEEE Int. Conf. Image Process. (ICIP)*, 2022, pp. 1581–1585.

[28] W. Zhang, R. R. Martin, and H. Liu, "A saliency dispersion measure for improving saliency-based image quality metrics," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 6, pp. 1462–1466, 2018.

[29] B. W. Tatler, "The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions," *J. Vis.*, vol. 7, no. 14, pp. 4–4, 2007.

[30] K. Seshadrinathan and A. C. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos," *IEEE Trans. Image Process.*, vol. 19, no. 2, pp. 335–350, 2009.

[31] I. Avcibas, B. Sankur, and K. Sayood, "Statistical evaluation of image quality measures," *J. Electron. Imaging*, vol. 11, no. 2, pp. 206–223, 2002.

[32] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004.

[33] Z. Wang, E. Simoncelli, and A. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, vol. 2, 2003, pp. 1398–1402 Vol.2.

[34] H. Sheikh and A. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, 2006.

[35] W. Xue, L. Zhang, X. Mou, and A. C. Bovik, "Gradient magnitude similarity deviation: A highly efficient perceptual image quality index," *IEEE Trans. Image Process.*, vol. 23, no. 2, pp. 684–695, 2014.

[36] C. G. Bampis, P. Gupta, R. Soundararajan, and A. C. Bovik, "Speed-qa: Spatial efficient entropic differencing for image and video quality," *IEEE Signal Process. Lett.*, vol. 24, no. 9, pp. 1333–1337, 2017.

[37] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, M. Manohara *et al.*, "Toward a practical perceptual video quality metric," *The Netflix Tech Blog*, vol. 6, no. 2, p. 2, 2016.

[38] S. Jain, P. Yarlagadda, S. Jyoti, S. Karthik, R. Subramanian, and V. Gandhi, "Vinet: Pushing the limits of visual modality for audio-visual saliency prediction," 2021.

[39] J. Antkowiak *et al.*, "Final report from the video quality experts group on the validation of objective models of video quality assessment march 2000," *Final Rep. Video Qual. Experts Group Valid. Obj. Models Video Qual. Assess. March 2000*, 2000.

[40] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, "Predicting Human Eye Fixations via an LSTM-based Saliency Attentive Model," *IEEE Trans. Image Process.*, vol. 27, no. 10, pp. 5142–5154, 2018.

[41] S. Jia and N. D. Bruce, "Eml-net: An expandable multi-layer network for saliency prediction," *Image Vis. Comput.*, vol. 95, p. 103887, 2020.

[42] J. Lou, H. Lin, D. Marshall, D. Saupe, and H. Liu, "Transalnet: Towards perceptually relevant visual saliency prediction," *Neurocomputing*, 2022.

[43] K. Min and J. J. Corso, "Tased-net: Temporally-aggregating spatial encoder-decoder network for video saliency detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 2394–2403.

[44] R. Droste, J. Jiao, and J. A. Noble, "Unified Image and Video Saliency Modeling," in *Proc. 16th Eur. Conf. Comput. Vis. (ECCV)*, 2020.

[45] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695–4708, 2012.

[46] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, 2013.

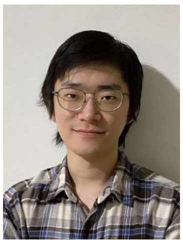
[47] X. Min, K. Gu, G. Zhai, J. Liu, X. Yang, and C. W. Chen, "Blind quality assessment based on pseudo-reference image," *IEEE Trans. Multimedia*, vol. 20, no. 8, pp. 2049–2062, 2018.

[48] X. Min, G. Zhai, K. Gu, Y. Liu, and X. Yang, "Blind image quality estimation via distortion aggravation," *IEEE Trans. Broadcast.*, vol. 64, no. 2, pp. 508–517, 2018.

[49] X. Min, Y. Gao, Y. Cao, G. Zhai, W. Zhang, H. Sun, and C. Chen, "Exploring rich subjective quality information for image quality assessment in the wild," *arXiv preprint arXiv:2409.05540*, 2024.

[50] W. Wang, J. Shen, J. Xie, M.-M. Cheng, H. Ling, and A. Borji, "Revisiting video saliency prediction in the deep learning era," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 220–237, 2021.

Xinbo Wu received his M.S. and Ph.D. degrees from the School of Computer Science and Informatics, Cardiff University, Cardiff, U.K., in 2020 and 2024. He has been the visiting scholar of Konstanz University, Konstanz, Germany. His research interests include visual quality assessment, visual perception and attention, and human-computer interaction.



Jianxun Lou received his B.Eng. degree from Central South University, Changsha, China, in 2018, and his M.S. and Ph.D. degrees from the School of Computer Science and Informatics at Cardiff University, Cardiff, U.K., in 2020 and 2024, respectively. He is currently a lecturer at Northeast Electric Power University in Jilin, China. His research interests include visual perception modelling and visual quality assessment.

Zhengyan Dong received the M.Sc. degree from the University of Birmingham, U.K., in 2019. He is currently pursuing the Ph.D. degree with the School of Computer Science and Informatics at Cardiff University. His research interests include machine learning, visual perception models, and visual quality assessment.

Fan Zhang received the B.Sc. and M.Sc. degrees from Shanghai Jiao Tong University, Shanghai, China, in 2005 and 2008, respectively, and the Ph.D. degree from the University of Bristol, Bristol, U.K., in 2012. He is currently a Senior Lecturer within the School of Computer Science, University of Bristol. He served as an Associate Editor for IEEE Transactions on Circuits and Systems for Video Technology (2022-2024), and was a guest editor of IEEE Journal on Emerging and Selected Topics in Circuits and Systems (in 2024) and Frontiers in Signal Processing (in 2022). Fan is also a member of the Visual Signal Processing and Communications Technical Committee associated with the IEEE Circuits and Systems Society. His research interests focus on low-level computer vision including video compression, quality assessment, super resolution and video frame interpolation.

Paul L. Rosin is a Professor at the School of Computer Science & Informatics, Cardiff University. Previous posts include Brunel University, Joint Research Centre, Italy and Curtin University of Technology, Australia. His research interests include computer vision, remote sensing, mesh processing, non-photorealistic rendering, performance evaluation, shape analysis, facial analysis, and cultural heritage.

Hantao Liu received the Ph.D. degree from the Delft University of Technology, Delft, The Netherlands in 2011. He is currently a Professor at the School of Computer Science and Informatics, Cardiff University, Cardiff, U.K.