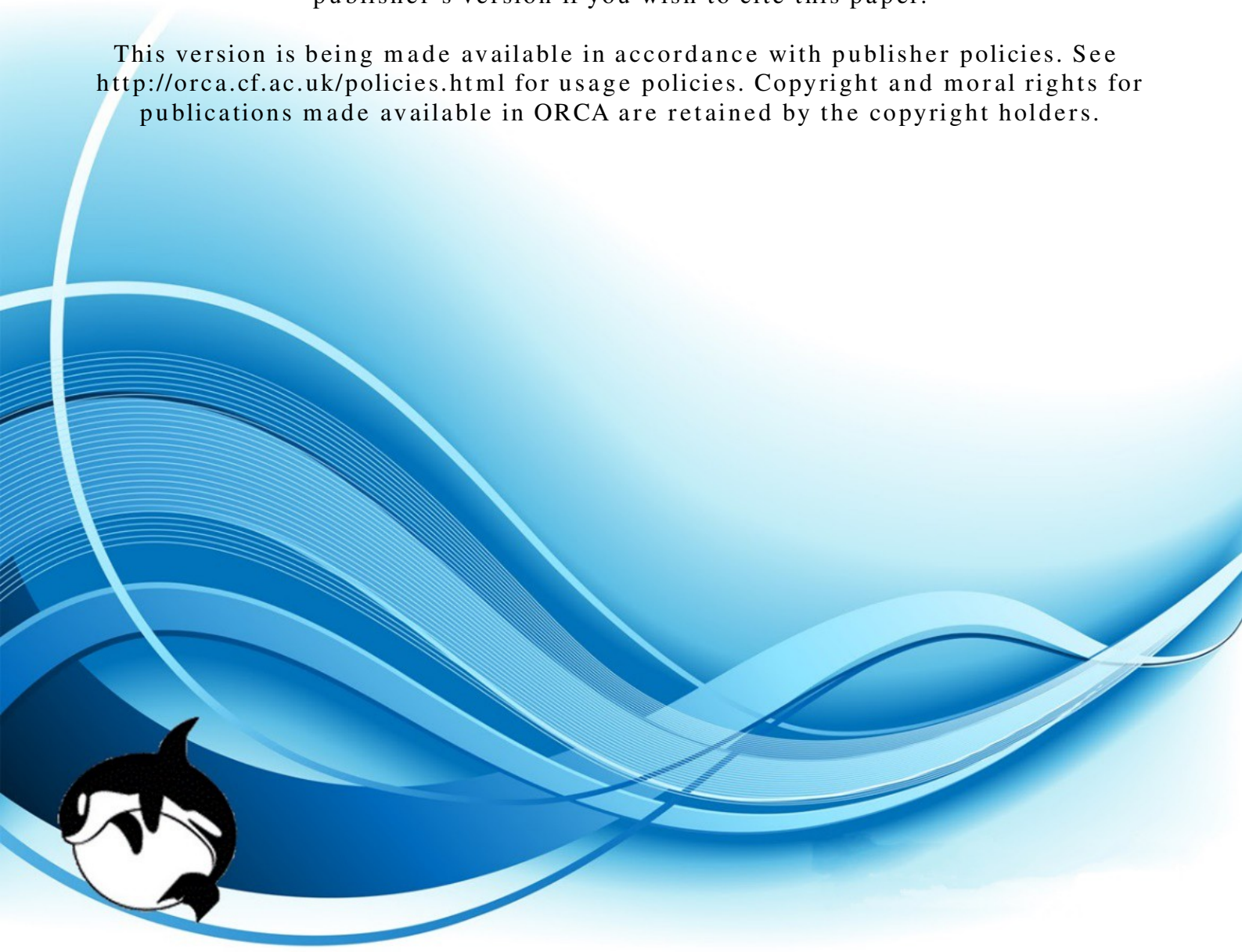


SSPNet: Predicting Visual Saliency Shifts

Huasheng Wang, Jianxun Lou, Xiaochang Liu, Hongchen Tan, Roger Whitaker, and Hantao Liu

***Abstract*—When images undergo quality degradation caused by editing, compression or transmission, their saliency tends to shift away from its original position. Saliency shifts indicate visual behaviour change and therefore contain vital information regarding perception of visual content and its distortions. Given a pristine image and its distorted format, we want to be able to detect saliency shifts induced by distortions. The resulting saliency shift map (SSM) can be used to identify the region and degree of visual distraction caused by distortions, and consequently to perceptually optimise image coding or enhancement algorithms. To this end, we first create a largest-of-its-kind eye-tracking database, comprising 60 pristine images and their associated 540 distorted formats viewed by 96 subjects. We then propose a computational model to predict the saliency shift map (SSM), utilising transformers and convolutional neural networks. Experimental results demonstrate that the proposed model is highly effective in detecting distortion-induced saliency shifts in natural images.**

***Index Terms*—Saliency, saliency shift, eye-tracking, transformer, convolutional neural networks.**

I. Introduction

SALIENCY prediction aims to mimic where humans look in a visual scene using computational technologies. Models of visual saliency are useful for many applications such as image compression [1]–[3], image quality assessment [4], [5], and salient object detection [6]–[10]. Eye movements under free-viewing conditions serve as the psychophysical foundations of saliency modelling [11]. In order to support the development of computational saliency models, a number of eye-tracking databases have been created. For example, MIT dataset [12] is one of the most widely used databases for visual saliency modelling, which consists of 1003 natural indoor and outdoor scenes freely viewed by 15 observers. CAT2000 [13] contains 4000 images from 20 different categories freely viewed by 24 observers. To make a large-scale dataset, SALICON [14] employed mouse clicks as a proxy for eye movement, generating a benchmark of 20,000 images.

Over the past few decades, notable progress has been made in computational modelling of visual saliency. Various saliency models have been produced [15]–[25]. In early stage research, the proposed approaches mainly adopt low-level hand-crafted visual features, including texture, colour and intensity [15]–[17]. Also, some heuristic saliency priors, e.g., texts and faces [26] are integrated into saliency models to augment their predictive ability. Due to the lack of higher-level visual features, these models cannot adequately reflect human visual attention. Recently, with the development of deep learning techniques, convolutional neural networks (CNN) have advanced the models for saliency prediction, leading to significantly improved results. The Ensembles of Deep Networks (eDN) [18] represents one of the deep learning-based models that utilise shallow CNN to predict saliency of natural images. The visual representations learned by CNN are more robust and comprehensive than hand-crafted features. Many deep learning-based saliency models have been proposed and achieved further success in visual saliency prediction. In the literature, AlexNet [27], VGGNet [28], and GoogleNet [29] have been applied on the pre-trained networks to learn features for saliency prediction. By comparing these three CNN architectures, Huang et al. [20] found that VGGNet is more suitable than the other two architectures for saliency prediction. Deep Visual Attention (DVA) [22] utilises one VGGNet-based encoder and three decoders to generate multi-scale feature representations for saliency detection. MSI-Net [23] employs VGGNet as the backbone and a skip architecture to extract multi-scale features that are fused by Atrous Spatial Pyramid Pooling [30]. To simulate explicit mechanisms of the human visual attention, Long short-term memory (LSTM) module is integrated into a saliency prediction model [21].

In many real-world applications, digital images are inevitably subject to visual distortions caused by e.g., compression, transmission or manipulation. Knowing viewers' behavioural responses to these distortions via eye-tracking would provide valuable information for the optimisation of modern imaging systems from the human-centric perspective. However, little is known about saliency of distorted images.

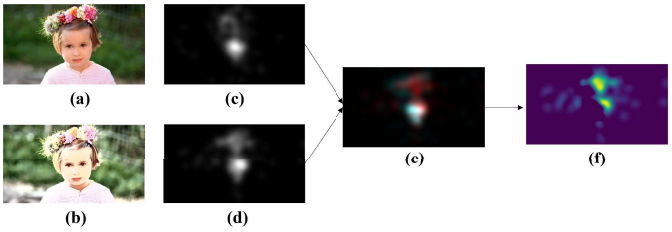


Fig. 1. Illustration of saliency shift map (SSM). (c) and (d) are saliency maps of the pristine image (a) and distorted image (b). (e) visualises the superimposition of (c) and (d), where the regions of visual saliency shifts are highlighted in red colour. (f) represents the SSM.

Huasheng Wang, Jianxun Lou, Roger Whitaker, and Hantao Liu are with the School of Computer Science and Informatics, Cardiff University, CF24 4AG Cardiff, United Kingdom.

Xiaochang Liu is with the School of Materials, Sun Yat-sen University, Guangzhou 510275, China.

Hongchen Tan is with the Institute of Artificial Intelligence, Beijing University of Technology, Beijing 100124, China

The work of Huasheng Wang was supported by China Scholarship Council under Grant 202106060056.

The work of Jianxun Lou was supported by China Scholarship Council under Grant 202008220129.

This work of Hongchen Tan was supported by National Natural Science Foundation of China 62201020, China Postdoctoral Science Foundation (BX20220025, 2021M700303), Beijing Postdoctoral Science Foundation 2022-ZZ-069, and Chaoyang Postdoctoral Science Foundation 2022ZZ-34.

Corresponding author: Jianxun Lou (louj2@cardiff.ac.uk)

A recent study has shown that when distortions occur in an image, its saliency tends to shift away from its original position, as illustrated in Fig. 1. Being able to detect visual saliency shifts (VSS) would benefit a variety of application scenarios. First, in critical visual tasks e.g., diagnostic imaging, the appearance of distortions in medical images impacts radiologists' diagnostic performance [31]. VSS can be used to measure the image readers' degree of distraction from the lesion detection task, and consequently to improve the diagnostic efficiency and accuracy. Second, in image compression and enhancement methods, performance optimisation can be often achieved through smart parameterisations in the regions of interest (ROI) and background regions (BR) [32]–[34], e.g., different levels of compression are applied to ROI and BR to reduce bitrate without compromising overall image quality. In such a scenario, VSS can be used to identify the BG that would be actually attended by viewers when distortions occur, and subsequently targeted parameterisation processes can be effectively applied to these affected image regions. It should be noted that the concept of "saliency shifts" is often used to describe the attention changes (or shift of attention [35]) in dynamic scenes. It refers to moving the focus of attention from one location to another, which can be driven intentionally by the use of visual cues or automatically by the abrupt onset of a stimulus [36], [37]. For example, in [38], the term "saliency shifts" is specifically used to indicate the phenomenon that salient object(s) may dynamically change in the video. In the context of image quality, the concept of "saliency shifts" refers to the re-allocation of attentional resources when distortions are introduced into a pristine image, and consequently indicating saliency that is shifted away from its original positions in the pristine image. Now, this raises a new research question, i.e., given a pristine image and its distorted format, how to detect distortion-induced saliency shifts. To make a model to predict the saliency shift map (SSM) as illustrated in Fig. 1, it requires groud truth representations of VSS via psychovisual experimentation and data. Then, the ultimate model aims to generate a topographic map that represents saliency of scene locations driven by distortions introduced to the pristine image.

II. Related work and contributions

Psychovisual studies have been undertaken to reveal the phenomenon of distortion-induced visual saliency shifts (VSS), demonstrating the significant difference in saliency between a pristine image and its distorted format. In [39] an eye-tracking study was conducted to prob the impact of distortions on the saliency of pristine images. It is found that visual distortions including white noise, blurring and compression artifacts significantly affect saliency patterns of pristine images. In [40], the study shows that distortions caused by JPEG compression can significantly change the saliency of pristine images; and the degree of saliency changes is found to be dependent on the level of compression. The eye-tracking study in [41] demonstrates that saliency patterns alter as visual distortions occur in a pristine image, and that the extent of saliency changes is related to the strength of distortion. These psychovisual studies provide empirical evidence that visual distortions cause can significant saliency shifts from its original places in the pristine images. Also, they suggest the importance of collecting eye-tracking data under free-viewing conditions for image quality research. This is to ensure the obtained saliency reflects the bottom-up, stimulus-driven attention rather than top-down, task-driven aspects of visual attention [42], [43]. However, it should be noted that these studies remain in an exploratory stage, representing a limited number of human subjects and a small degree of stimulus variability. A further critical issue for these studies is that their eye-tracking data is strongly biased due to the involvement of intensive stimulus repetition where observers learnt to detect visual artifacts in viewing of the same natural scene content (with multiple variations of distortion) repeatedly. In this case, the recorded eye-tracking data is potentially contaminated due to strong carry-over effects, and therefore cannot be used as the ground truth to study the real interactions between pristine scene saliency and unnatural visual distortions.

In a recent study [44], a refined experimental methodology is proposed to enable a reliable collection of eye-tracking data for pristine images and their altered formats containing distortions of various types and strength levels. This methodology applies dedicated control mechanisms to eliminate subject bias due the involvement of stimulus repetition; and provide experimental conditions and requirements for achieving saturated/stable ground truth saliency data. By using this methodology, a highly reliable eye-tracking database SIQ288 including 288 images distorted with different types of distortion at various degradation levels was created [44]. An exhaustive statistical analysis was performed to demonstrate the significance of saliency changes caused by the addition of distortions. Although the SIQ288 database represents the best-of-its-kind in the literature, it only contains 18 pristine images which might limit its use for computational saliency modelling where the diversity in natural scene content plays a critical role. This implies the need of a new benchmark which encompasses sufficient image content diversity.

There are two main contributions we want to make in this paper. First, we follow the experimental methodology in [44] to create a new eye-tracking database specifically for visual saliency shifts (VSS). The test images are taken from the CUID database [45], which represents a large degree of stimulus variability in terms of the amount and diversity of natural scene content, as well as a systematic simulation of different types and levels of distortions. The new eye-tracking study results in a largest-of-its-kind ground truth VSS dataset, comprising 600 visual stimuli of 10 different categories (with multiple variations of distortion) and eye movement recordings of 96 participants. The ground truth VSS manifests itself as a function of natural scene category and visual distortion. Second, we propose a novel model based on deep convolutional neural network to predict the saliency shift map (SSM) induced by distortion, namely saliency shift prediction network (SSPNet). The SSPNet is built on an end-to-end framework where saliency prediction of the pristine image and that of the distorted image are jointly optimised.

III. Eye-tracking Study

The eye-tracking study aims to generate the ground truth data for visual saliency shifts (VSS), reflecting a sufficient degree of stimulus variability in terms of natural scene category, as well as the type and level of distortion. More importantly, the well-thought-out experimental design devised in [44] is adopted in this study to ensure the validity and reliability of the resulting eye-tracking data. Details of the study are described below.

A. Stimuli

The stimuli were taken from the CUID database [45], which included 60 high-quality pristine images (1920×1080 pixels) of 10 natural scene categories including ACT (Action), BNW (Black and White), CGI (Computer-Generated Imagery), IND (Indoor), OBJ (Object), ODM (Outdoor Manmade), ODN (Outdoor Natural), PAT (Pattern), POT (Portrait), and SOC (Social), as illustrated in Fig. 2. In the CUID database, each pristine image was degraded with three distortion types including contrast change (i.e., CC), JPEG compression (i.e., JPEG), and motion blur (i.e., MB) at three distortion levels. In stimulating distortions, distortion parameters of each distortion type were set/adjusted via visual inspection performed by image quality experts. This process was to ensure that the distorted images created from each stimulus (per distortion type) reflected three distinct levels of perceived quality: Q1 (representing perceptible but not annoying artifacts), Q2 (representing noticeable and annoying artifacts), and Q3 (representing very annoying artifacts). The details of implementation can be found in [45]. This gives a total of 600 test stimuli including originals.

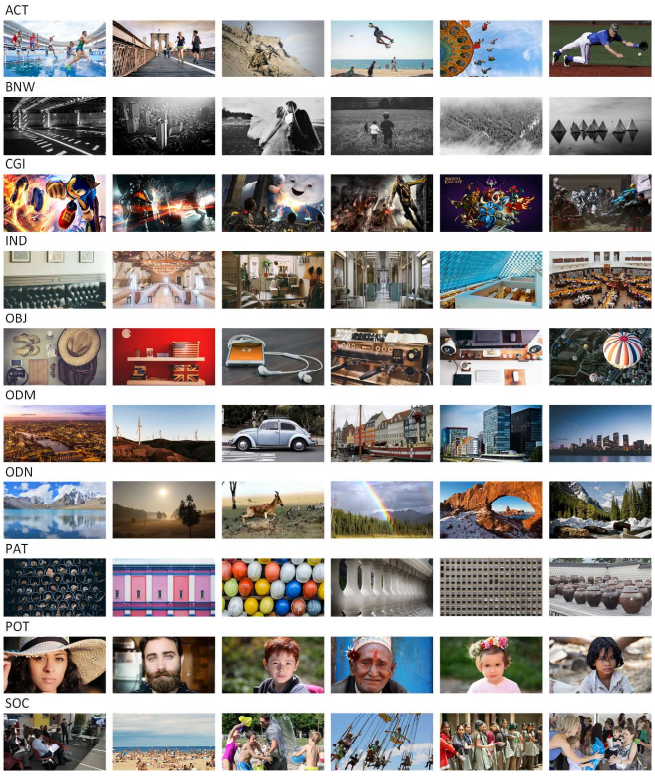


Fig. 2. Illustration of the 60 source images of 10 distinct categories used in our study. The categories are ACT (Action), BNW (Black and White), CGI (Computer-Generated Imagery), IND (Indoor), OBJ (Object), ODM (Outdoor Manmade), ODN (Outdoor Natural), PAT (Pattern), POT (Portrait), and SOC (Social).

B. Eye-tracking experiment

In our experiment, each pristine image is associated with nine distorted images of the same scene content. The existence of stimulus repetition poses significant challenges for eye-tracking [44], leading to subject bias caused by carry-over effects such as fatigue, boredom and learning from practice and experience. To eliminate the bias we employ the experimental methodology devised in [44] in our eye-tracking study.

Following the protocol of [44] for a between-subjects experiment, we divided the set of stimuli into six partitions of 100 images each. In each partition, we only included a maximum of two repeated formats of the same scene content. A total of 96 subjects were recruited to partake in the experiment, being 48 males and 48 females with ages ranging from 19 to 55 years. The subjects were divided into six groups of 16 subjects each (with 8 males and 8 females); and each group was randomly assigned to one of the partitions of stimuli so that each subject only had to complete one session of viewing 100 images. This provided a sample size of 16 subjects per test stimulus, which has been proven sufficient for generating a reliable saliency map [44]. To minimise carry-over effects, we also divided each session per subject into two sub-sessions with a "washout" period of 5 hours in between, which actually allowed each subject to view 50 images (i.e., half partition of stimuli) without stimulus repetition in a separate sub-session. More specifically, we applied the following control mechanisms for each partition of stimuli in our experiment: (1) half of the subjects viewed the first half of stimuli first, and half of the subjects viewed the second half first; (2) the stimuli in each sub-session were presented to each subject in a random order; (3) a mixture of all distortion types and the full range of distortion levels was contained in each sub-session.

The eye-tracking experiment was conducted in the Visual Computing laboratory at Cardiff University in a standard office environment set up as per the International Telecommunication Union (ITU) standards [46]. The laboratory represented a fully controlled viewing environment to ensure consistent experimental conditions, i.e., low surface reflectance and approximately constant ambient light. The test stimuli were displayed on a 19-inch LCD screen, with a native resolution of 1920×1080 pixels. The viewing distance was maintained around 60cm. The subjects' eye movements were recorded using a non-invasive SensoMotoric Instrument (SMI) Red-m advanced eye tracking device at a sampling rate of 250 Hz. Prior to the start of the actual experiment, each subject was given written instructions about the testing procedure. The subjects were instructed to view the stimuli in a natural way "view the image as you normally would". Each image was presented for five seconds followed by a mid-gray screen of two seconds.

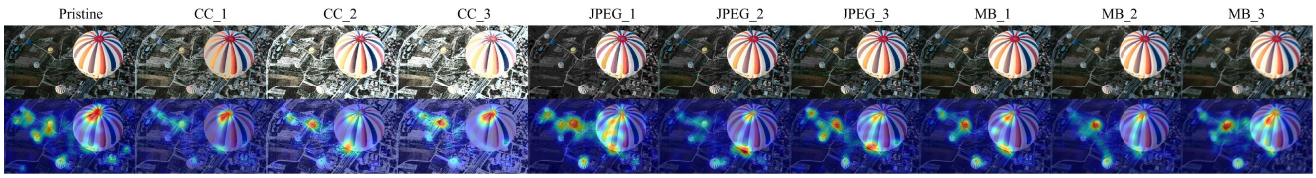


Fig. 3. Illustration of saliency maps for one of the natural scenes including the pristine content and various formats of distorted content. The top row shows the stimuli. The label of the distorted image is expressed as $type\_level$ indicating the type and intensity level of distortion respectively. The type of distortion includes CC (contrast change), JPEG (JPEG compression) and MB (motion blur); and the intensity level of 1, 2, and 3 denotes low, medium, and high distortion, respectively. The bottom row shows the corresponding saliency maps.

C. Ground truth of visual saliency shifts

By use of the experiment protocol of [44], a saturated/stable saliency map is produced via the stringent control of stimulus presentation and requirement of subject sample size. By converting the resulting data of our eye-tracking experiment to a saliency map for each image, this guarantees the saliency (as defined in the image quality context [44]) represents stimulus-driven, bottom-up attention for an average human observer [47]–[49]. To this end, fixations are extracted from the raw eye movement data using the SMI BeGaze Analysis Software. Note a fixation is rigorously defined by SMI's Software using the dispersal and duration based algorithm established in [50]. To construct a saliency map each fixation (out of all fixations over all 16 subjects) location gives rise to a gray-scale patch that stimulates the foveal vision of the human visual system. The activity of the patch manifests a Gaussian distribution, with its width approximating the size of the fovea (i.e., 2°of visual angle converting to 45 pixels width in our study). The saliency map (SM) is calculated as:

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

where $\mathrm{SM}(x,y)$ represents the saliency map; $(x_i,\ y_i)$ and N represent the spatial coordinates of i-th fixation and the total number of fixations, respectively; σ is the standard deviation of the Gaussian ($\sigma = 45$ pixels in our study), and the method of its determination can be found in detail in [51]. Fig. 3 illustrates the saliency maps for one of the natural scenes including the pristine content and various formats of distorted content.

Once saliency maps are generated, the ground truth of visual saliency shifts (VSS) is rendered. More specially, a saliency shift map (SSM) is created as follows:

$$\mathcal{SSM} = (\mathcal{SM}_d - \mathcal{SM}_r)_+, \quad (2)$$

where $\mathcal{SM}_d$ and $\mathcal{SM}_r$ represent the saliency maps for the distorted image and pristine image, respectively; and $()_+$ denotes the operation of taking the positive values in the matrix. As shwon in Fig. 1, $\mathcal{SSM}$ represents where people's attention is shifted when viewing an distorted image in comparison to the pristine image. As per the reliability validation in [44], saturation occurs in a saliency map with 12-16 participants when following the designed protocol, meaning the saliency map reaches the point at which no new information is observed. Furthermore, the protocol used also eliminates subject bias due to stimulus/content repetition. As a result, the visual saliency shifts (VSS) faithfully reflect the re-allocation of attentional resources when distortions are introduced into a pristine image. This means the observed differences in saliency between the pristine and distorted images (i.e., SSM) are from the differences in the image properties rather than from the subject and/or experiment bias. The saliency shift maps (SSMs) of the 540 distorted images constitute a new ground truth dataset for VSS, namely CUID-VSS database.

TABLE I
RESULTS OF THE ANOVA TO EVALUATE THE IMPACT OF DISTORTION TYPE, DISTORTION LEVEL AND IMAGE CONTENT DIVERSITY ON THE MEASURED VISUAL SALIENCY SHIFTS (VSS) USING CC, SIM OR KLD. "DF" DENOTES DEGREE OF FREEDOM, AND "SIG" DENOTES THE SIGNIFICANCE LEVEL.

ANOVA		CC	SIM	KLD
Source	df	Sig	Sig	Sig
Distortion type	2	0.04	0.03	<0.001
Distortion level	2	<0.001	<0.001	0.03
Image content diversity	9	<0.001	<0.001	0.002

D. Statistical analysis

To verify the merits of the new CUID-VSS database, we perform a statistical analysis on the observed tendencies in the changes of VSS included by the changes of image properties. To this end, hypothesis testing is conducted to evaluate the impact of three categorical variables including distortion type, distortion level and image content diversity on the visual saliency shifts (VSS). VSS can be quantified by a distribution-based saliency fidelity/similarity measure [52] namely Pearson's Correlation Coefficient (CC), Similarity (SIM) or Kullback-Leibler Divergence (KLD). An analysis of variance (ANOVA) is performed with the measured VSS as the dependent variable (note the test for the assumption of normality indicates that the dependent variable is normally distributed), and the distortion type, distortion level and image content diversity as independent variables. Table I summarises the results of the ANOVA, and shows that all main effects are statistically significant. Especially, as mentioned in Section II, the image content diversity is a critical new feature of the CUID-VSS database, which has been proven statistically significant ($P < 0.05$ at 95% level) in impacting the VSS. Overall, the statistical analysis demonstrates the statistical validity of the CUID-VSS database in supporting the computational modelling of the VSS.

IV. Proposed Saliency Shift Prediction Model

Our goal is to predict the saliency shift map (SSM) given a pristine image and a distorted format of the same scene content. The predicted SSM aims to be in close agreement with the ground truth saliency shifts. To solve this problem, we hereby propose a computational model based on deep learning named *Saliency Shift Prediction Network* (SSPNet). The overall architecture of SSPNet is shown in Fig. 4, which contains one encoder and two decoders to achieve the saliency shift prediction task. The input of SSPNet is composed of two images, i.e., the pristine image and its associated distorted image. The output is the saliency shift map (SSM). The details of SSPNet are presented below.

A. Architecture

In the encoder phase, a deep CNN-based backbone network (i.e., ResNet [53] or VGGNet [28]) is used as the saliency feature extractor in our model. Because each of these backbone network contains a variety of convolution layers, the SSPNet can benefit from the features of the deeper layers that encode more high-level semantics and of the shallower layers that carry richer low-level details. In order to fit the architecture to a saliency prediction task, similar to previous saliency models [21], [22], [24], all the fully-connected layers and the last pooling layer are discarded in the basic feature extraction phase. In our implementation, the CNN encoder of SSPNet is pre-trained on ImageNet [54], which is a widely used method of feature extraction for visual saliency prediction [21], [23]–[25], [55], [56]. In addition, we use the transformer technique in our model, due to its demonstrable effectiveness enhancing saliency prediction [57]. As can be seen from Fig. 4, the $\mathcal{F}_1$, $\mathcal{F}_2$, and $\mathcal{F}_3$ are the feature maps obtained from the end of the third to last, second to last, and final blocks of the encoder, respectively. Their spatial dimensions are $\frac{w}{8} \times \frac{h}{8}$, $\frac{w}{16} \times \frac{h}{16}$, $\frac{w}{32} \times \frac{h}{32}$. And then a 1 × 1 convolution layer is applied to reduce the computational burden, so that the channels of these feature maps are 512, 768, and 768, which is an established method in [46] for balancing computational cost and performance efficacy. The parameters of the encoder and transformer modules are shared by both of the decoders in our model.

With regard to the decoder phase, we employ two decoders with the same structure to predict two "hidden" saliency maps, one from the pristine image and one from the distorted image as the output of the hidden layer. These two "hidden" saliency maps are generated based on feature maps $\mathcal{F}_{1t}$, $\mathcal{F}_{2t}$, and $\mathcal{F}_{3t}$ from the transformer encoders. These feature maps are fused into the decoders by skip-connection and element-wise production (also known as Hadamard product), leading to multi-scale context-enhanced feature maps [57]. block_4 consists of two upscaling blocks and an additional sequence of operations using as the final transformation. Each upscaling block encompasses a 3×3 convolution, batch normalisation, ReLU activation, and an upsampling operation. After these two blocks, it proceeds with another sequence of operations including a convolution-batch normalisation-ReLU chain, followed by an additional convolution and ending with a sigmoid activation. The design of block_4 aims to transform the feature maps into a saliency map that matches the size of the input image. Finally, the two "hidden" saliency maps are combined using (2) to generate the output saliency shift map (SSM), where the three elements in (2) are constrained by respective loss functions (will be detailed below). Since the input image is 32-scale down-scaled by the encoder network, 2-scale up-samplings that adopt nearest-neighbor interpolation are performed to the feature maps in each decoder to obtain a output saliency map of the same size as the input. By adopting a dual-decoder architecture, the SSM can be generated directly from the model, which allows the loss function to constrain the SSM generation to yield optimal quantitative results. The advantages of using the dual-decoder are detailed in Section V-C.

B. Loss function

In modelling visual saliency using deep learning, loss functions are of fundamental importance [58]. In the literature of saliency prediction [21], [23]–[25], [57], many loss functions have been proven effective in improving the performance of saliency models. In this paper, we employ a linear combination of three metrics to form a loss function in our model, including Kullback-Leibler Divergence (KLD), Linear Pearson's Correlation Coefficient (CC), and Similarity (SIM). As mentioned above, three individual loss functions must be used to constrain the generation of the saliency map of the pristine image, the saliency map of the distorted image, and the saliency shift map, respectively, we now give the details of these loss functions. For the pristine image, we denote y_r and $\hat{y_r}$ as the predicted saliency map and the ground truth, and i indicates the ith pixel of y_r and $\hat{y_r}$. The loss function is defined as:

$$L_r(y_r, \hat{y_r}) = \lambda_1 L_{KLD}(y_r, \hat{y_r}) + \lambda_2 L_{CC}(y_r, \hat{y_r}) + \lambda_3 L_{SIM}(y_r, \hat{y_r}), \tag{3}$$

where $\lambda_1, \lambda_2, \lambda_3$ are the weights, and

$$L_{KLD}(y_r, \hat{y_r}) = \sum_i \hat{y_{r_i}} \log(\frac{\hat{y_{r_i}}}{\epsilon + y_{r_i}} + \epsilon), \tag{4}$$

where ϵ is a regularization constant and set to 1×10^{-8} ;

$$L_{CC}(y_r, \hat{y_r}) = \frac{cov(y_r, \hat{y_r})}{\sigma(y_r)\sigma(\hat{y_r})}, \tag{5}$$

where $cov(\cdot)$ represents covariance and $\sigma(\cdot)$ represents standard deviation;

$$L_{SIM}(y_r, \hat{y_r}) = \sum_i \min(y_{r_i}, \hat{y_{r_i}}). \tag{6}$$

Note, y_r and $\hat{y_r}$ are normalised so that $\sum_i y_{r_i} = \sum_i \hat{y_{r_i}} = 1$.

Similarity, for the distorted image, we denote y_d and $\hat{y_d}$ as the predicted saliency map and the ground truth. Therefore, the loss function is defined as:

$$L_d(y_d, \hat{y_d}) = \lambda_1 L_{KLD}(y_d, \hat{y_d}) + \lambda_2 L_{CC}(y_d, \hat{y_d}) + \lambda_3 L_{SIM}(y_d, \hat{y_d}), \tag{7}$$

where these weights are set to be the same as (3).

Finally, for the saliency shift map, we denote y_s and $\hat{y_s}$ as the predicted map and the ground truth, and L_s can be

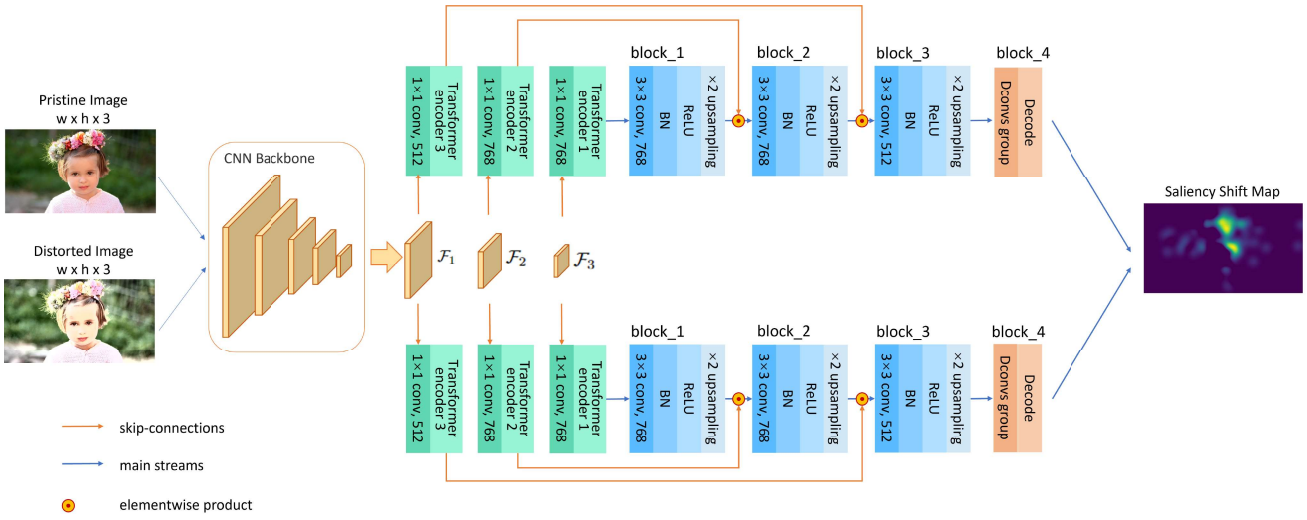


Fig. 4. Schematic overview of SSPNet. The proposed model contains one encoder and two decoders to predict the saliency shift map. The input of SSPNet is composed of a pair of pristine image and its associated distorted image, which is first processed by the CNN and transformer encoders, and then processed by different decoders. The saliency shift map is generated based on the outputs of the two decoders.

similarly defined as above. In addition, we add the square root loss function (L_2) to L_s to make the results smoother:

$$L_s(y_s, \hat{y_s}) = \alpha L_d(y_s, \hat{y_s}) + \beta L_2(y_s, \hat{y_s}). \quad (8)$$

Hence, our total loss function is defined as follows,

$$L_{total} = \gamma_1 L_d(y_d, \hat{y_d}) + \gamma_2 L_r(y_r, \hat{y_r}) + \gamma_3 L_s(y_s, \hat{y_s}). \quad (9)$$

C. Backward propagation process

Since the ground truth saliency shift map (SSM) is obtained by subtracting the reference from the distorted and then taking the positive values of the resulting matrix as seen in (2), it is worth discussing whether the model can be back-propagated to optimise the prediction. Piecewise functions are only not derivative at the breakpoints. For our model, the truncation function will be employed on the procedure that generates the SSM based on the saliency map of distorted image and that of the reference (i.e., the last procedure of the model). As a result of the above operation, the features that carry the information of image quality alteration will propagate back to update the parameters, and other irrelevant features will not cause the model to update the parameters. Generally, the truncation function is to facilitate the generation of SSM, and it is derivative per se. Using this method instead of directly subtracting the two saliency maps can avoid the problem that the negative values of the matrix are forcibly converted into 255 pixel values (white points) due to the unit8 encoding of images. Note the saliency shift map computed by (2) is able to accurately characterise the offsets of human visual attention as the image quality alters.

V. Experimental Results

In this section, we conduct a series of experiments to validate the proposed SSPNet model in predicting the visual saliency shifts. The details are given below.

A. Evaluation metrics

Various evaluation metrics have been proposed to assess the performance of saliency models. Generally, these metrics can be classified as location-based and distribution-based metrics depending on how the ground truth is represented [52]; the former adopts the fixation map (i.e., in the form of a binary image) and the latter uses the saliency map (i.e., in the form of a gray-scale image) as the ground truth for visual saliency evaluation. Previous studies have suggested that the selection of metrics used to evaluate saliency prediction results should depend on the task in hand [52]. Because the ground truth (i.e., saliency shift map) of our task is a density map, we must use distribution-based metrics including KLD, CC, and SIM to quantify the general performance of saliency models in this study. Details of these metrics can be found in [52] and in Section IV-B. For KLD, the closer the value is to zero, the better the agreement between prediction and ground truth. For the other two metrics, CC and SIM, higher values indicate higher performance.

B. Experimental setup

We divide the CUID-VSS database into training set and test set according to the ratio of 9:1. More specifically, the training set contains 540 images, including 54 pristine images and their associated 486 distorted images; and the test set contains 60 images, including 6 pristine images and their associated 54 distorted images. It is worth noting that we split the dataset randomly. The advantage of doing this is that it can increase the generalisation ability of our model.

To reduce the computational cost, all input images are resized and padded to a same size of 384×288 pixels, i.e., the dimensions of both the reference and distorted images are $384 \times 288 \times 3$ $(w \times h \times c)$. The network was implemented with the PyTorch framework on a single NVIDIA RTX 3060 GPU. Following a similar procedure in the state-of-the-art [21], [22],

[24], our model is initialised by the weights pre-trained on ImageNet [54] to reduce the risk of overfitting. Based on the pre-trained ResNet-50 backbone, our network features an end-to-end model for the whole training stage. For training, the Adam algorithm [59] is used to minimise the value of loss function. The learning rate is set to 5×10^{-5}, which is then multiplied by 0.1 for every 10 epochs. Models are trained with a batch size of 4 for 50 epochs with a stop patience of 5 epochs. At each stop patience, the model that performs best on the three evaluation metrics (i.e., CC, SIM, and KLD) is saved and then is used in the test phase. With regard to the hyperparameters $\lambda_1, \lambda_2, \lambda_3, \alpha, \beta, \gamma_1, \gamma_2, \gamma_3$, we set them to 5, -2, -1, 2, 0.5, 0.5, 0.5, 1, respectively base on our empirical experiments. The approach taken to determine these hyperparameters consists of two main steps including initial assignment and fine-tuning. More specifically, for the visual saliency loss (i.e., L_r and L_d), $\lambda_1, \lambda_2, \lambda_3$ are initially assigned as different levels of importance to the sub-metrics/sub-losses to balance their contribution to the final loss as the sub-losses' values use different scales. Also, as a lower value of KLD (a higher value of CC or SIM) represents a higher agreement between predicted saliency and ground truth, λ_1 should intuitively be set to negative and λ_2, λ_3 should be positive. Based on above, the assignment of these hyperparameters can be initially determined, as the same principles applied in previous studies [21], [25], [57]. Then, we perform a "grid search"-like method, i.e., fine-tuning individual weights by adjusting one weight and fixing the remaining weights to optimise the performance on the validation set [60]. The goal is to find a combination of weights that allows the model to achieve good and balanced scores on the saliency evaluation metrics. Similarly, the set of weights $\gamma_1, \gamma_2, \gamma_3$, and the set of weights α, β are determined based on the same approach of initial assignment combined with grid search. It should be noted that hyperparameter optimisation is a challenging task for a deep-learning model, and there are no formal rules for hyperparameter tuning [60]. We follow the conventional process to search the space of possible decisions and find the best-performing ones empirically.

Since k-fold cross-validation is a robust measure to prevent overfitting, providing a more reliable estimate of a model's performance on completely unseen data [57], [60], we implement k-fold ($k = 6$) cross-validation in our experiments. More specifically, we divide the dataset into six non-overlapping folds, each containing 90 distorted images (originated from 10 source scenes as shown in Fig. 2). To ensure that there is no data leakage, no shared source scenes between folds is allowed – each fold contains scenes extracted from one of the columns of the 10×6 grid gallery as shown in Fig. 2. In the 6-fold cross-validation process, one fold is used as the test set, one fold as the validation set, and the remaining four folds as the training set. This process is repeated six times, with each of the six folds used exactly once as the test set. The six results are then averaged to produce a single estimation for the model's generalisation performance.

TABLE II
PERFORMANCE COMPARISON OF DIFFERENT MODEL DESIGN CONCEPTS: SINGLE-DECODER VERSUS DUAL-DECODER

Model Design	CUID-VSS		
	CC ↑	SIM ↑	KLD ↓
Single-decoder	0.332	0.341	13.576
Dual-decoder	0.739	0.623	0.764

TABLE III
PERFORMANCE COMPARISON OF MODEL VARIANTS USING DIFFERENT ENCODER BACKBONE NETWORKS

Encoder Backbone	CUID-VSS		
	CC ↑	SIM ↑	KLD ↓
VGG (SSPNet)	0.692	0.577	0.856
ResNet-18 (SSPNet)	0.721	0.589	0.799
ResNet-50 (SSPNet)	0.739	0.623	0.764

C. Ablation study

Now, we want to verify the design of our proposed model. To quantify the relative importance of different key components of the model, we conduct a comprehensive ablation study. First, it is critical to justify the necessity for using a dual-decoder architecture for predicting saliency shifts, which is the core design concept of our proposal. Without the dual-decoder, the alternative is to use a single decoder to predict the saliency of the pristine image and that of the distorted image separately, and to generate the saliency shift map (SSM) using (2). We would argue that our proposed dual-decoder design gives better predictive performance due to its back-propagation ability as described in Section IV-C. To verify this, we conduct experiments to analyse the effectiveness of single-decoder design versus dual-decoder design in predicting the visual saliency shifts. Table II illustrates the performance comparison of the single-decoder design versus dual-decoder design. It demonstrates that our proposed model design outperforms the single-decoder design. There are two benefits for our end-to-end approach towards predicting the SSM. First, our approach is more efficient than the single-decoder method, because it can directly generate the final SSM instead of synthesising the SSM using the output of the network. Second, during training, the loss function is used to constrain the generation of the SSM, so that the results are automatically optimised. However, the single-decoder method inevitably suffers from the superposition of prediction errors, i.e., errors of predicting the saliency map of the pristine image and errors of predicting the saliency map of the distorted image. In addition, we also demonstrate the performance comparison of using different encoder backbone networks, i.e., VGG, ResNet-18, and ResNet-50, as the results shown Table III. According to previous studies [61], CC and SIM are the best metrics for measuring the distortion-induced saliency variation. Therefore, we rely on CC and SIM to assess different backbones. By comparing VGG, ResNet-18, and ResNet-50, it is found that ResNet-50, which has higher capability of representation, can provide on average better performance. Therefore, ResNet-50 is adopted as the backbone encoder in the implementation of this study to construct the SSPNet. Transformers have been proven useful in augmenting saliency prediction [57]. Since saliency prediction

TABLE IV
PERFORMANCE COMPARISON OF SSPNET WITH VERSUS WITHOUT TRANSFORMERS. **BOLD FONT** INDICATES THE BEST PERFORMANCE SCORE.

Model Name	CUID-VSS		
	CC ↑	SIM ↑	KLD ↓
SSPNet (without transformers)	0.575	0.497	6.724
SSPNet (with transformers)	**0.739**	**0.623**	**0.764**

TABLE V
PERFORMANCE COMPARISON OF OUR PROPOSED SSPNET WITH INTUITIVE APPROACH (IA) APPROACH USING STATE-OF-THE-ART SALIENCY MODELS ON THE CUID-VSS DATABASE. **BOLD FONT** INDICATES THE BEST PERFORMANCE SCORE.

Model Name	CUID-VSS		
	CC ↑	SIM ↑	KLD ↓
IA_EML-NET [24]	0.299	0.296	13.954
IA_SAM-VGG [21]	0.344	0.315	12.865
IA_SAM-ResNet [21]	0.365	0.319	12.681
IA_UNISAL [25]	0.396	0.378	12.105
IA_MSI-Net [23]	0.405	0.392	11.241
SSPNet (Ours)	**0.739**	**0.623**	**0.764**

forms the key component of the SSPNet architecture, adding transformers is naturally expected to benefit the overall model performance. In additional, the long-range representational capabilities of transformers may help characterise the saliency-specific differences and correspondences between the pristine and distorted images for the SSPNet. To verify the added value of transformers in the proposed architecture, we perform experiments to compare the performance of SSPNet with versus without transformers. As the results illustrated in Table IV, the contribution of transformers to the SSPNet is rather significant.

D. Comparative experiments: SSPNet versus Intuitive approach (IA)

In this paper, we have taken a new approach to building a dedicated model for the prediction of saliency shifts. Intuitively, one way to achieve saliency shifts is to separately predict the saliency maps of the pristine and distorted images, and generate the saliency shifts by comparing these two resulting saliency maps. However, this approach is prone to multiple sources of error, i.e., one source of error from the saliency prediction of the pristine image and the other from the saliency prediction of the distorted image. Combing the results of two erroneous prediction tasks will further deteriorate the final estimation of saliency shifts. The rationale behind our approach is to directly predict a saliency shift map (SSM) via learning a dual-decoder network, and consequently to minimise the overall error for the prediction task. This approach takes into account the saliency-specific differences and correspondences between the pristine and distorted images, resulting in a more accurate estimation of the saliency shifts.

To quantitatively verify the contribution of the proposed approach taken in our model, we conduct comparative experiments using state-of-the-art saliency prediction models (i.e., as per the MIT saliency benchmark [62]). First, we apply the intuitive approach (IA) where a single-decoder model is used to predict saliency once for the pristine image and once for the distorted image, and then the SSM is generated by taking the difference of the two saliency maps using equation (2). For fairness, all IA models were fine-tuned (using the same training strategy as described in Section V.B) on the CUID-VSS dataset (note 60 pristine images and 540 distorted images for respective fine-tuning tasks) to achieve the best possible performance on the task of VSS prediction. Table V lists the performance of the IA approach using different saliency models, compared with the performance of the proposed SSPNet. Fig. 5 provides the visual comparison of the performance of these models. It can be seen from the table and figure that the IA approach generally fails in predicting saliency shifts, which implies the need of a dedicated approach towards this specific application. Our proposed SSPNet can sufficiently capture the distortion-induced saliency shifts in natural scenes, demonstrating the effectiveness of the proposed solution.

TABLE VI
PERFORMANCE COMPARISON OF OUR PROPOSED SSPNET WITH MODEL COMPETITORS (MC) BASED ON STATE-OF-THE-ART SALIENCY MODELS ON THE CUID-VSS DATABASE. **BOLD FONT** INDICATES THE BEST PERFORMANCE SCORE.

Model Name	CUID-VSS		
	CC ↑	SIM ↑	KLD ↓
MC_EML-NET [24]	0.286	0.296	13.385
MC_SAM-VGG [21]	0.568	0.501	7.664
MC_SAM-ResNet [21]	0.581	0.509	6.548
MC_UNISAL [25]	0.641	0.553	1.915
MC_MSI-NET [23]	0.708	0.601	1.253
SSPNet (Ours)	**0.739**	**0.623**	**0.764**

E. Comparative experiments: SSPNet versus model competitors (MC)

In the SSPNet, a new dual-decoder architecture is proposed to solve the specific problem of saliency shift prediction. To the best of our knowledge, there is no existing model so far in the literature that dedicates to the visual saliency shifts (VSS). However, the saliency prediction component embedded in the architecture of the SSPNet plays a crucial role in achieving the VSS prediction task. Therefore, to critically evaluate the proposed SSPNet, we conduct further comparative experiments by implementing model competitors (MC). To rigorously create a model competitor, we replace the saliency prediction component in the proposed SSPNet using a state-of-the-art saliency model (i.e., as per the MIT saliency benchmark [62]) and adapting it to the dual-decoder architecture. For fairness, all MC alternatives were fine-tuned (using the same training strategy as described in Section V.B) on the CUID-VSS dataset to achieve the best possible performance on the task of VSS prediction. Also, all models were implemented using the same coding/experimental environment including settings of learning rate, optimizer and other relevant parameters. Table VI lists the performance of the MC based on different saliency models, compared with the performance of the proposed SSPNet. Fig. 6 provides the visual comparison of the performance of these models. It can be seen from the table and figure that our proposed SSPNet significantly outperforms other model competitors in predicting saliency shifts.

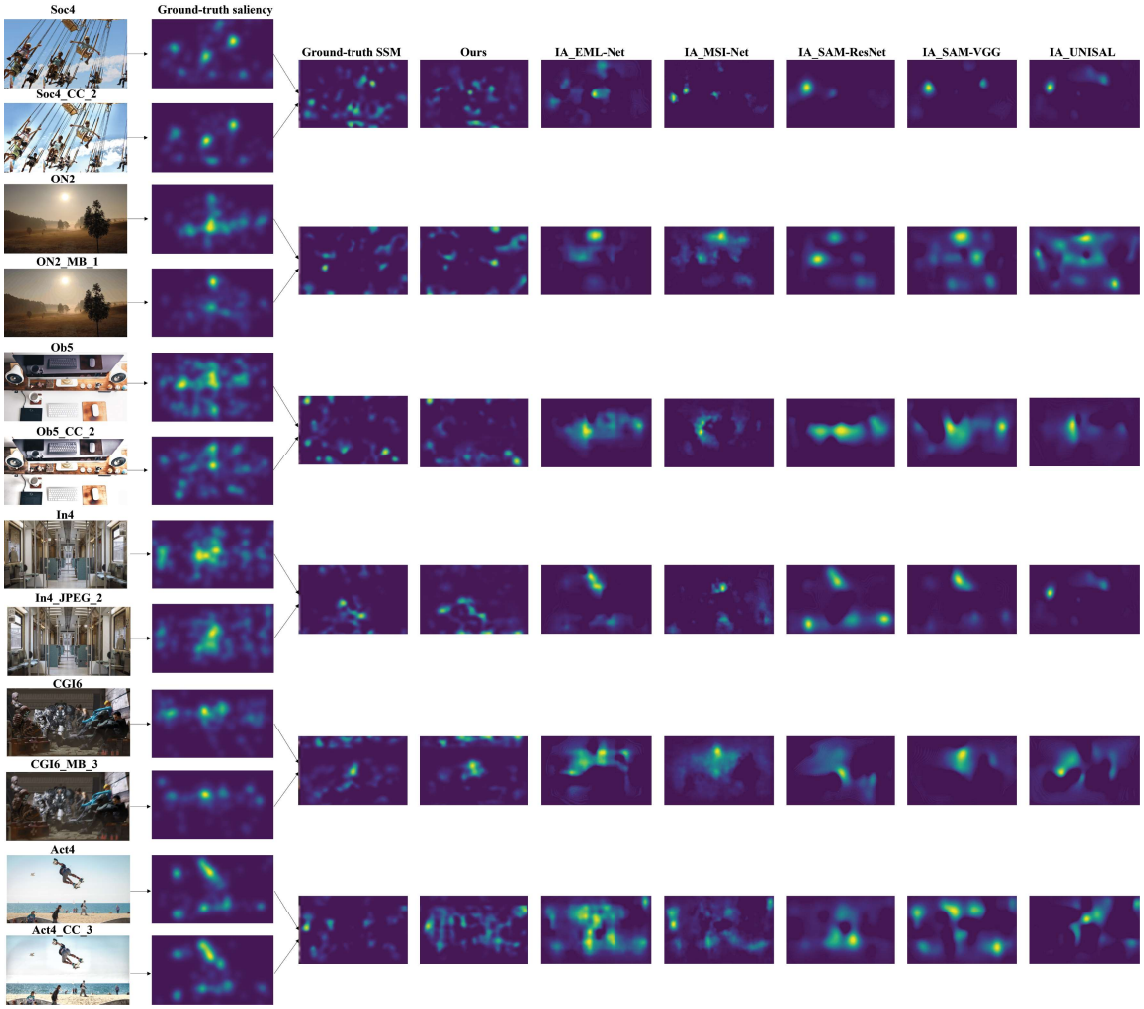


Fig. 5. Visualisations of comparative experiments: SSPNet versus intuitive approach (IA). From left to right, the columns represent pristine/distorted images, ground truth saliency, ground truth saliency shift map (SSM), and the prediction outputs of our model and five models using the intuitive approach (IA) based on state-of-the-art saliency models, respectively.

TABLE VII
RESULTS OF STATISTICAL SIGNIFICANCE TESTING FOR MODEL PERFORMANCE QUALIFIED BY CC, SIM OR KLD. "1" MEANS THAT THE DIFFERENCE IN PERFORMANCE IS STATISTICALLY SIGNIFICANT (P<0.05 AT THE 95% CONFIDENCE LEVEL). "0" MEANS THAT THE DIFFERENCE IN PERFORMANCE IS NOT STATISTICALLY SIGNIFICANT.

	MC_EML-NET	MC_SAM-VGG	MC_SAM-ResNet	MC_UNISAL	MC_MSI-NET
	CC-SIM-KLD	CC-SIM-KLD	CC-SIM-KLD	CC-SIM-KLD	CC-SIM-KLD
SSPNet(ours)	1-1-1	1-1-1	1-1-1	1-1-1	1-1-1

To verify whether the results of model performance as listed in Table VI are statistically significant, hypothesis testing is conducted using the statistical methods in [44]. The significance evaluation is based on 540 VSS outputs (note inputs of 540 distorted stimuli originated from 60 originals) of all testing results in a 6-fold cross-validation for each of the models. Therefore, each model produces 540 data points for a performance measure calculated by CC, SIM, or KLD between the ground truth VSS and predicted VSS. Now, we compare the performance between two models using their CC, SIM or KLD data points (540 each). When two samples in question are both normally distributed, an independent samples t-test is performed; otherwise, in the case of non-normality, a non-parametric version (i.e., Mann-Whitney U test) analogy to an independent samples t-test is performed. The results of the statistical evaluation are listed in Table VII. This means our proposed model is statistically significantly (P<0.05 at the 95% confidence level) better than every other model in predicting the visual saliency shifts.

VI. DISCUSSION

The selection of backbone networks has been demonstrated an important approach to improving the performance of deep-learning models [63], [64]. In particular, recent studies [57], [65] show that using backbones with a superior representational capability could further boost a saliency prediction

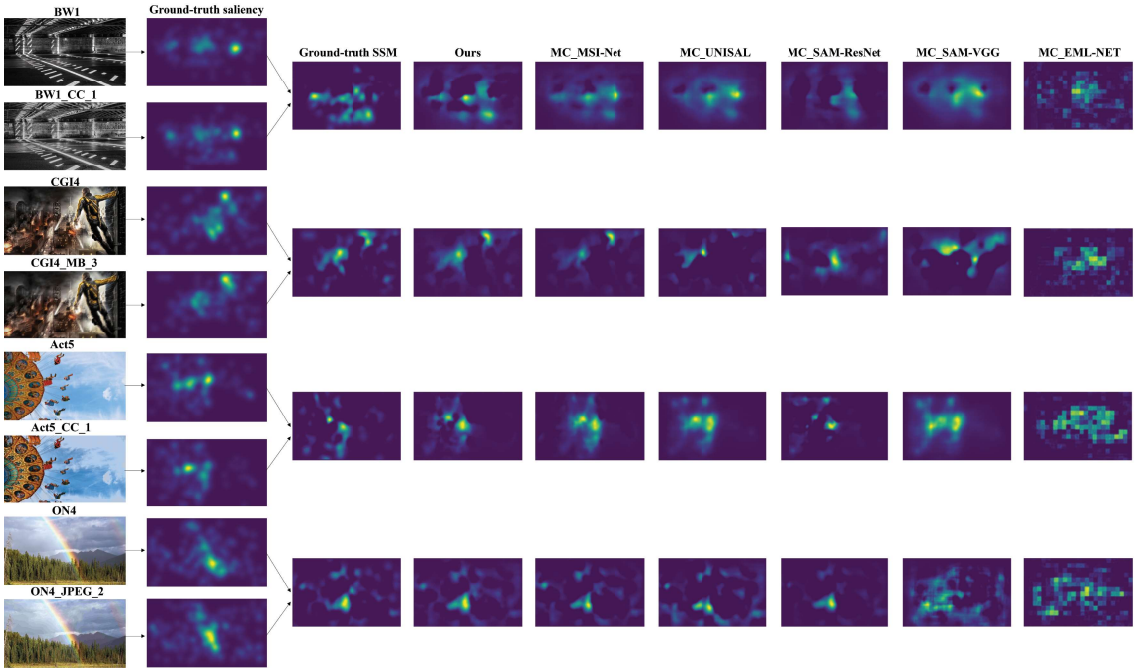


Fig. 6. Visualisations of comparative experiments: SSPNet versus model competitors (MC). From left to right, the columns represent pristine/distorted images, ground truth saliency, ground truth saliency shift map (SSM), and the prediction outputs of our model and five model competitors (MC) based on state-of-the-art saliency models, respectively.

model's performance for both general and specific applications. The same trend holds for saliency shift prediction in this paper as evidenced by the ablation study for VGG, ResNet-18 and ResNet-50 in Section V.C. While comparing the added value of different backbones to the performance of a deep-learning model is useful, other methods could be sought to compare their representational ability as alternative ways to assess the contribution of different backbones. For example, methods for visualising high-dimensional data, e.g., t-SNE [66] could be applied to compare different backbones in terms of separability of features in the latent space. We have conducted a preliminary experiment for SSPNet on one test set from our k-fold cross-validation using t-SNE. We took these 90 test images, extracted the features of $\mathcal{F}_1$, $\mathcal{F}_2$, and $\mathcal{F}_3$ from our SSPNet architecture and then used t-SNE to compute a 2-dimensional embedding that respects the high-dimensional distances [58], [66]. This was done three times each using VGG, ResNet-18 or ResNet-50 as the backbone. By visualising the embeddings where images are displayed at their embedded locations, we found that ResNet-50 indeed shows the best discriminative capability (with ResNet-18 being the second best).

In this paper, we limit the loss functions for the saliency prediction component of SSPNet to the category of "saliency-inspired" loss functions. It should, however, be noted that the choice of loss functions plays a significant role in modelling visual saliency [58]. In [58], the study provides a comprehensive analysis of the impact of the use of four different categories of loss functions (including saliency-inspired and non-saliency-inspired) and use of six linear combinations of different losses on saliency prediction. It shows that a careful design of the loss function can significantly improve the performance of a saliency prediction model. We would expect that improving the predictive power of the saliency prediction component will enhance the overall performance of the saliency shift prediction of SSPNet. Although designing a dedicated and robust loss function is outside of the scope of this paper, this will be treated in a separate contribution in the future work.

VII. Conclusion

In this paper, we have presented our work towards predicting visual saliency shifts (VSS) induced by distortions. To tackle this new research problem, we first carried out an eye-tracking experiment to gain ground truth of VSS, resulting in a new CUID-VSS database. Then, we devised a computational model SSPNet to predict VSS using deep learning. By integrating transformers into CNNs, our proposed SSPNet model significantly benefits from capturing long-range spatial information at multiple perceptual levels, leading to state-of-the-art performance in predicting VSS. The software and database will be made publicly available to facilitate further research on visual saliency shifts.

References

[1] C. Guo and L. Zhang, "A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression," *IEEE Transactions on Image Processing*, vol. 19, no. 1, pp. 185–198, 2010.

[2] S. Li, M. Xu, Y. Ren, and Z. Wang, "Closed-form optimization on saliency-guided image compression for hevc-msp," *IEEE Transactions on Multimedia*, vol. 20, no. 1, pp. 155–170, 2018.

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

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

[5] S. Yang, Q. Jiang, W. Lin, and Y. Wang, "SGDNet: An End-to-End Saliency-Guided Deep Neural Network for No-Reference Image Quality Assessment," in *Proceedings of the 27th ACM International Conference on Multimedia*. New York, NY, USA: Association for Computing Machinery, 2019.

[6] L. Ye, Z. Liu, L. Li, L. Shen, C. Bai, and Y. Wang, "Salient object segmentation via effective integration of saliency and objectness," *IEEE Transactions on Multimedia*, vol. 19, no. 8, pp. 1742–1756, 2017.

[7] R. Shi, Z. Liu, H. Du, X. Zhang, and L. Shen, "Region diversity maximization for salient object detection," *IEEE Signal Processing Letters*, vol. 19, no. 4, pp. 215–218, 2012.

[8] B. Alexe, T. Deselaers, and V. Ferrari, "What is an object?" *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 73–80, 2010.

[9] N. Imamoglu, W. Lin, and Y. Fang, "A saliency detection model using low-level features based on wavelet transform," *IEEE Transactions on Multimedia*, vol. 15, no. 1, pp. 96–105, 2013.

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

[11] L. Itti and C. Koch, "Computational modelling of visual attention," *Nature reviews neuroscience*, vol. 2, no. 3, pp. 194–203, 2001.

[12] T. Judd, F. Durand, and A. Torralba, "A benchmark of computational models of saliency to predict human fixations," MIT Computer Science and Artificial Intelligence Lab (CSAIL), Cambridge, MA, USA, Tech. Rep. MIT-CSAIL-TR-2012-001, 01 2012.

[13] A. Borji and L. Itti, "CAT2000: A large scale fixation dataset for boosting saliency research," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2015 workshop on "Future of Datasets"*, 2015, arXiv preprint arXiv:1505.03581.

[14] M. Jiang, S. Huang, J. Duan, and Q. Zhao, "SALICON: Saliency in context," in *IEEE Conf. on Comput. Vis. and Pattern Recognit.*, 2015, pp. 1072–1080.

[15] D. Walther and C. Koch, "Modeling attention to salient proto-objects," *Neural Networks*, vol. 19, no. 9, pp. 1395–1407, 2006.

[16] J. Harel, C. Koch, and P. Perona, "Graph-Based Visual Saliency," in *Proceedings of the 19th International Conference on Neural Information Processing Systems (NIPS)*, ser. NIPS'06. Cambridge, MA, USA: MIT Press, 2006, p. 545–552.

[17] E. Erdem and A. Erdem, "Visual saliency estimation by nonlinearly integrating features using region covariances," *Journal of Vision*, vol. 13, no. 4, pp. 11–11, 03 2013.

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

[19] S. Yang, G. Lin, Q. Jiang, and W. Lin, "A Dilated Inception Network for Visual Saliency Prediction," *IEEE Transactions on Multimedia*, vol. 22, no. 8, pp. 2163–2176, 2020.

[20] X. Huang, C. Shen, X. Boix, and Q. Zhao, "SALICON: Reducing the semantic gap in saliency prediction by adapting deep neural networks," in *IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 262–270.

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

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

[23] A. Kroner, M. Senden, K. Driessens, and R. Goebel, "Contextual encoder–decoder network for visual saliency prediction," *Neural Networks*, vol. 129, pp. 261–270, 2020.

[24] S. Jia and N. D. Bruce, "EML-NET: An expandable multi-layer network for saliency prediction," *Image and Vision Computing*, vol. 95, p. 103887, 2020.

[25] R. Droste, J. Jiao, and J. A. Noble, "Unified image and video saliency modeling," in *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part V*, ser. Lecture Notes in Computer Science, A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, Eds., vol. 12350. Springer, 2020, pp. 419–435.

[26] M. Cerf, E. P. Frady, and C. Koch, "Faces and text attract gaze independent of the task: Experimental data and computer model," *Journal of Vision*, vol. 9, no. 12, pp. 10–10, 11 2009.

[27] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of The ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[28] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *2015 International Conference on Learning Representations (ICLR)*, 2015.

[29] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1–9.

[30] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2018.

[31] H. Liu, J. Koonen, M. Fuderer, and I. Heynderickx, "The relative impact of ghosting and noise on the perceived quality of mr images," *IEEE Transactions on Image Processing*, vol. 25, no. 7, pp. 3087–3098, 2016.

[32] Y. Patel, S. Appalaraju, and R. Manmatha, "Saliency driven perceptual image compression," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 227–236.

[33] S. Li, M. Xu, Y. Ren, and Z. Wang, "Closed-form optimization on saliency-guided image compression for hevc-msp," *IEEE Transactions on Multimedia*, vol. 20, no. 1, pp. 155–170, 2017.

[34] H. Alers, J. Redi, H. Liu, and I. Heynderickx, "Studying the effect of optimizing image quality in salient regions at the expense of background content," *Journal of Electronic Imaging*, vol. 22, no. 4, pp. 043 012–043 012, 2013.

[35] A. Johnson and R. W. Proctor, *Attention: Theory and practice*. Sage, 2004.

[36] C. Koch and S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry." *Human neurobiology*, vol. 4, no. 4, pp. 219–227, 1985.

[37] M. S. Peterson, A. F. Kramer, and D. E. Irwin, "Covert shifts of attention precede involuntary eye movements," *Perception & psychophysics*, vol. 66, no. 3, pp. 398–405, 2004.

[38] D.-P. Fan, W. Wang, M.-M. Cheng, and J. Shen, "Shifting more attention to video salient object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 8554–8564.

[39] 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 Southwest Symposium on Image Analysis and Interpretation*. IEEE, 2008, pp. 73–76.

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

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

[42] E. Niebur, C. Koch, and W. Rucklidge, "Computational architectures for attention. the attentive brain," *R Parasuraman, R., ed.,. MIT Press, Cambridge, Massachusetts*, pp. 163–186, 1998.

[43] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 185–207, 2012.

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

[45] L. Lévêque, J. Yang, X. Yang, P. Guo, K. Dasalla, L. Li, Y. Wu, and H. Liu, "CUID: A New Study Of Perceived Image Quality And Its Subjective Assessment," in *2020 IEEE International Conference on Image Processing (ICIP)*, 2020, pp. 116–120.

[46] R. I.-R. BT, "Methodology for the subjective assessment of the quality of television pictures," *International Telecommunication Union*, 2002.

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

[48] H. Liu and I. Heynderickx, "Visual attention in objective image quality assessment: Based on eye-tracking data," *IEEE transactions on Circuits and Systems for Video Technology*, vol. 21, no. 7, pp. 971–982, 2011.

[49] U. Engelke, H. Kaprykowsky, H.-J. Zepernick, and P. Ndjiki-Nya, "Visual attention in quality assessment," *IEEE Signal Processing Magazine*, vol. 28, no. 6, pp. 50–59, 2011.

[50] D. D. Salvucci and J. H. Goldberg, "Identifying fixations and saccades in eye-tracking protocols," in *Proceedings of the 2000 Symposium on Eye Tracking Research & Applications*, ser. ETRA '00. New York, NY, USA: Association for Computing Machinery, 2000, p. 71–78.

[51] C. Privitera and L. Stark, "Algorithms for defining visual regions-of-interest: comparison with eye fixations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 9, pp. 970–982, 2000.

[52] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "What do different evaluation metrics tell us about saliency models?" *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 3, pp. 740–757, 2019.

[53] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[54] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[55] M. Kümmerer, L. Theis, and M. Bethge, "Deep Gaze I: Boosting saliency prediction with feature maps trained on ImageNet," *arXiv preprint arXiv:1411.1045*, 2014.

[56] M. Kummerer, T. S. A. Wallis, L. A. Gatys, and M. Bethge, "Understanding low- and high-level contributions to fixation prediction," in *IEEE International Conference on Computer Vision (ICCV)*, Oct 2017, pp. 4799–4808.

[57] J. Lou, H. Lin, D. Marshall, D. Saupe, and H. Liu, "TranSalNet: Towards perceptually relevant visual saliency prediction," *Neurocomputing*, vol. 494, pp. 455–467, 2022.

[58] A. Bruckert, H. R. Tavakoli, Z. Liu, M. Christie, and O. Le Meur, "Deep saliency models : The quest for the loss function," *Neurocomputing*, vol. 453, pp. 693–704, 2021.

[59] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015.

[60] F. Chollet, "Deep learning with python," 2017.

[61] X. Yang, F. Li, and H. Liu, "A measurement for distortion induced saliency variation in natural images," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–14, 2021.

[62] M. Kümmerer, Z. Bylinskii, T. Judd, A. Borji, L. Itti, F. Durand, A. Oliva, and A. Torralba, "MIT/Tübingen Saliency Benchmark," https://saliency.tuebingen.ai/.

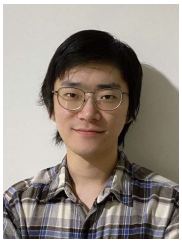
[63] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 11 966–11 976.

[64] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 9992–10 002.

[65] J. Lou, H. Lin, P. Young, R. White, Z. Yang, S. Shelmerdine, D. Marshall, E. Spezi, M. Palombo, and H. Liu, "Predicting radiologists' gaze with computational saliency models in mammogram reading," *IEEE Transactions on Multimedia*, pp. 1–14, 2023.

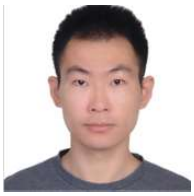
[66] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

Huasheng Wang received the B.Eng. from Xiangtan University, in 2018 and the M.S. degree from Dalian University of Technology in 2021. He is now pursuing his Ph.D. degree at the School of Computer Science and Informatics, Cardiff University, Cardiff, UK. His interests are Saliency Prediction, Image Quality Assessment and Depth Estimation.



Jianxun Lou received the B.Eng. from Central South University, Changsha, China, in 2018 and the M.S. degree from Cardiff University, Cardiff, UK, in 2020. He is now pursuing his Ph.D. degree at the School of Computer Science and Informatics, Cardiff University, Cardiff, UK.

Xiaochang Liu is currently pursuing the bachelor's degree within School of Materials at Sun Yat-sen University, China. Her research interests include mathematical modelling and data analytics.



Hongchen Tan is a Lecturer of Artificial Intelligence Research Institute at Beijing University of Technology. He received Ph.D degrees in computational mathematics from the Dalian University of Technology in 2021. His research interests are Person Re-identification, Image Synthesis, and Referring Segmentation.

Roger Whitaker is a Professor at the School of Computer Science and Informatics, Cardiff University, Cardiff, U.K. His research concerns the intersection of machine and human intelligence, including human behaviour. He is an Area Editor for Online Social Networks and Media (Elsevier) and Associate Editor for Social Network Analysis and Mining (Springer).

Hantao Liu received the Ph.D. degree from the Delft University of Technology, Delft, The Netherlands in 2011. He is currently an Associate Professor with the School of Computer Science and Informatics, Cardiff University, Cardiff, U.K. He is an Associate Editor of IEEE Transactions on Circuits and Systems for Video Technology and IEEE Signal Processing Letters.