Deep Visual Saliency on Stereoscopic Images

Anh-Duc Nguyen, Jongyoo Kim, Member, IEEE, Heeseok Oh, Haksub Kim, Weisi Lin, Fellow, IEEE, and Sanghoon Lee, Senior Member, IEEE

Abstract—Visual saliency on stereoscopic 3D (S3D) images has been shown to be heavily influenced by image quality. Hence, this dependency is an important factor in image quality prediction, image restoration and discomfort reduction, but it is still very difficult to predict such a nonlinear relation in images. In addition, most algorithms specialized in detecting visual saliency on pristine images may unsurprisingly fail when facing distorted images. In this paper, we investigate a deep learning scheme named Deep Visual Saliency (DeepVS) to achieve a more accurate and reliable saliency predictor even in the presence of distortions. Since visual saliency is influenced by low-level features (contrast, luminance and depth information) from a psychophysical point of view, we propose seven low-level features derived from S3D image pairs and utilize them in the context of deep learning to detect visual attention adaptively to human perception. During analysis, it turns out that the low-level features play a role to extract distortion and saliency information. To construct saliency predictors, we weight and model the human visual saliency through two different network architectures, a regression and a fully convolutional neural networks (CNNs). Our results from thorough experiments confirm that the predicted saliency maps are up to 70% correlated with human gaze patterns, which emphasize the need for the hand-crafted features as input to deep neural networks in S3D saliency detection.

Index Terms—Saliency prediction, stereoscopic image, distorted image, convolutional neural network, deep learning.

I. INTRODUCTION

Visual attention is an important characteristic in the human visual system (HVS). In general, humans can locate and fixate on any region of their interest in a scene, at a glance, without the need to scan the whole scene. This region usually possesses some unique characteristics which make it stand out and attractive to humans and is termed as salient region, region of interest or saliency. Saliency detection has long been an important topic in image processing for its vast engineering applications such as image compression/cropping/thumbnailing, tracking/surveillance, or image rendering. Numerous studies have been proposed to predict salient regions on natural 2D scenes [16]–[18, 21, 22, 51] and the performance is getting more accurate and reliable. However, with the introduction of disparity in S3D images, these models fail to capture the saliency driven by depth perception, which may result in predictions far different from where humans actually look. Despite the rising popularity of S3D contents, there are only a few models proposed to patch this shortcoming [4, 9, 10, 38, 52]. In addition, another important cue that may influence visual attention is image quality. In the 2D case, several works on the influence of image quality to human attention have been done and it is observed that most common distortions do not significantly affect human gaze pattern [42, 43]. In S3D case, however, it has been shown that commonly known distortions strongly affect visual attention [26, 32] because distortions greatly affect depth perception of humans [39], which is a major factor driving human gaze. For a concrete example, when an image is blurry, viewers tend to shift their gaze towards near objects [26]. If in an image, the objects of interest are already near, then the shift in attention is little. Conversely, if the objects of interest in an image is far and small, it is difficult to recognize such objects, so instead they shift their gazes towards the foreground objects. In such case, the difference in gazing behaviors when viewing pristine and distorted images is very significant. An illustration of such phenomenon is shown in Fig. 1.

Another decisive factor contributing to the performance of a predictor is the computational resource. Previously, many studies follow a bottom-up principle; i.e. hand-engineered
filters and features are developed to predict saliency. These methods usually predict salient objects well without the need for too much resource. However, as pointed out in [25], the saliency maps produced by these studies are usually not well-correlated to human ground truth data. Hence, many top-down methods [25,55], which rely on machine learning models, have been proposed to benefit from human ground truth. Recently, the performance of saliency detectors has been hugely boosted thanks to the introduction of deep learning, which in most cases is deep neural networks. One disadvantage of this approach is the model complexity. Deep networks usually require a huge amount of resource and data for training. Such data and resource may not be readily available so a more resource-friendly approach is preferred in many situations. One idea is developing several simple hand-engineered features in order to reduce the depth of deep models.

In this paper, in order to reduce the model complexity of deep learning without sacrificing accuracy, we propose a deep visual saliency (DeepVS) framework which uses a variety of features suitable for the problem of saliency detection as deep networks’ input. To better present which factors and features we consider, the major ones are listed as follows.

**Distortion:** In [19,26], it has been shown how distortion affects human visual attention. Fig. 1 demonstrates how the saliency maps of blurred images are different from those of the pristine images. Moreover, distortion in 3D vision also damages the depth perception and results in many problems such as binocular rivalry and binocular suppression [32,39,45]. These problems affect the visual attention more significantly than in the 2D case.

**Depth/Disparity:** 3D perception has been shown to greatly drive human gaze [9,24,38]. This can be confirmed by the fact that the main theme of a picture tends to be captured so that it appears to be in the center and foreground of the image. In the 2D image case, depth estimation from a single image is a very ill-posed and difficult problem. Reconstructing the depth of a scene is equivalent to reconstructing the 3D space from the 2D image space, which requires to recover the whole null space of the projection matrix from the singularity. This difficulty is one of the reasons why this important factor is mostly overlooked in literature. In S3D case, however, we can utilize the overlap from two images of the same scene and backtrack to the real life scene to recover the relative depth effortlessly. Since we want to model the attention of humans, this feature is of the utmost importance in our study.

**Content characteristics:** Low-level features such as color, luminance, edges are well-known factors driving human attention. In deep learning literature, however, these features are being overlooked and only raw RGB (red, green, and blue) images are considered instead. This is convenient in many problems where we do not have much prior knowledge. In the problem of saliency prediction, we have a very strong prior about which features and characteristics of an image induce human attention. It has been shown that handcrafted features can boost the performance of deep learning further than raw inputs [13]. Also, it is widely accepted that several first layers of the deep networks usually learn simple and low-level features. Hence, explicitly using low-level features as input is reasonable and reduces the model complexity as well.

**Relevant HVS property:** When viewing S3D contents, it is observed that viewers are attracted to different regions in an image compared with viewing that image in 2D. Moreover, when visual contents are distorted, the fusability of S3D pairs is affected, which results in phenomena such as binocular rivalry or binocular suppression. Thus, it is necessary to quantify visual saliency based on the optical and physiological characteristics of humans including binocular fusion and foveation [27,33].

Although these factors have been discussed individually and partially in previous research, there has been no study taking account of all the above factors. As tabulated in Table I, the major drawback of previous work is the lack of a multi-aspect look at visual saliency. Some advanced features like distortion, depth, and relevant HVS property are frequently omitted. Nevertheless, considering different factors altogether is very important because it plays an important role in image quality, image restoration, and discomfort reduction. How different factors affect saliency and how each layer in a deep network contributes to the prediction of salient regions has not yet been conducted thoroughly. It has been long-known that color, luminance, and depth/disparity greatly influence saliency but most studies only employ them as features without quantifying how much their influences are. Also, deep networks have been frequently used in saliency detection but no attempt has been made to understand how semantically each layer aligns with the final prediction. Such studies are necessary because they can give us an insight into selection of features and deep network architecture in order to enhance the performance quality.

Our proposed framework, visually described in Fig. 2,

<table>
<thead>
<tr>
<th>Research</th>
<th>Considered factors</th>
<th>HVS</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>[16]–[18,21,22,23]</td>
<td>–</td>
<td>–</td>
<td>v</td>
</tr>
<tr>
<td>[9,10,38,52]</td>
<td>–</td>
<td>v</td>
<td>v</td>
</tr>
<tr>
<td>[19,26]</td>
<td>v</td>
<td>–</td>
<td>v</td>
</tr>
<tr>
<td>[27,33]</td>
<td>v</td>
<td>v</td>
<td>v</td>
</tr>
<tr>
<td>DeepVS</td>
<td>v</td>
<td>v</td>
<td>v</td>
</tr>
</tbody>
</table>

### TABLE I

**Considered factors in saliency detection literature. 'SHALLOW' refers to models that are bottom-up or handcrafted-feature machine learning based while "DEEP" indicates deep learning-based methods.**

[16]–[18,21,22,23] | – | – | v | – | – | – | shallow |
| [9,10,38,52] | – | v | v | – | – | – | shallow |
| [19,26] | v | – | v | v | v | – | – |
| [27,33] | v | v | v | v | v | – | deep |
| DeepVS | v | v | v | v | v | – | deep |
consists of two stages: feature extraction and learning. In the feature extraction stage, inspired by [26], from an S3D pair, we extract the representative features, which are derived from color, luminance and disparity information, to detect human gaze patterns with respect to psychophysical characteristics in the HVS. For the learning stage, we choose a deep learning approach, which is known to outperform conventional machine learning methods given sufficient data, to validate the features. We propose to use a regression convolutional network (DeepVS-R) and a fully convolutional network (DeepVS-F) on the extracted features to prove that any architecture can benefit from our hand-crafted features. We first compare the results from the model against those from current state-of-the-art models in saliency detection, and then analyze how performance can be improved with bells and whistles. In addition, we thoroughly analyze the importance of each feature to visual attention. Finally, we visualize what the hidden layers of the network actually learn. Through rigorous experiments and benchmarks, we show that our saliency detection scheme is competitive with currently top-notch models while maintaining a simple and straightforward architecture. In our view, there are three important contributions here:

1) We introduce different low-level information useful for indicating saliency in S3D images even under presence of distortion. To demonstrate the significance of the features, we resort to different deep CNN architectures and show performance gains against state-of-the-art models in saliency detection as well as the cases in which only RGB input is considered.

2) We analyze the importance of the feature maps through experiments and thereby demonstrate the need for the hand-crafted features as inputs to deep networks in S3D saliency detection.

3) We propose several modifications to the conventional convolutional neural networks (CNNs) to better enhance the performance of the saliency detection model. Some analyses on the hidden layers of the fully convolutional network in terms of semantic meaning are also presented.

II. FEATURE EXTRACTION

A. Feature maps

Attention behaviors are strongly affected by low-level components such as luminance and color gradients, disparity and depth discontinuity [9,10,26,38]. Luminance and color usually define the uniqueness and interestingness of an object while human viewers are prone to concentrate on nearer objects, which is simulated by depth/disparity. A notable characteristic of these features is that they help the deep network extract semantically meaningful and human-level understandable information in hidden layers, which we demonstrate later in Section IV-D. Therefore, in this study, the data fed to the deep learning model is a set of features extracted from color, luminance and disparity. The procedure to extract these maps is simple and similar to data augmentation, feature normalization, and zero-phase component analysis whitening, which are popular pre-processing techniques in the deep learning area, in many ways. The detailed process is described below.

1) Binocular information. To model the reliable visual attention on S3D images, a virtually synthesized image resembling what humans perceive in accordance with binocular vision system is required rather than just a left or right one. Perceptually, each S3D image pair is fused in the brain as a virtual single image, called a cyclopean image. The artificial cyclopean image used in this study is a composition of a stereo image pair, the corresponding disparity map and Gabor filter responses (4 orientations: horizontal, both diagonals and vertical at a frequency of 3.67 cycles/degree) [6]. This form of cyclopean image has been verified to correlate strongly with the fused S3D image in our brain [6]. Let I be a cyclopean image. The expression of I is

\[ I(x, y) = W_L(x, y) I_L(x, y) + W_R(x + d, y) I_R(x + d, y) \]  

where \( I_L \) and \( I_R \) are the left and right images, \( W_L \) and \( W_R \) are magnitudes of the Gabor filter responses, \((x, y)\) indicates pixel coordinate and \(d\) is the disparity corresponding to each pixel. The RGB channels of \( I \) are the first three of the seven maps.

2) Content information. We use the luminance and color gradients [10,17,27], which reveal the change of luminance and color in spatial domain. The normalized luminance and color gradient maps are extracted from the cyclopean images. Let \( \Delta_l \) denote the luminance gradient. The normalized luminance gradient is derived as

\[ \Delta_l = \frac{1}{\delta_{l}^{M}} \sqrt{(\nabla_x I)^2 + (\nabla_y I)^2} \]  

where \( \nabla_x I \) and \( \nabla_y I \) are the spatial gradients of the cyclopean image with respect to the horizontal and vertical directions, respectively, and \( \delta_{l}^{M} \) is the maximum value of the square root term.

To capture the color information, the cyclopean images are converted from RGB to CIELab color space to become \( I_{Lab} \), the image intensity in \( Lab \) scale. Similar to the normalized luminance gradient, the normalized color gradient \( \Delta_c \) is defined as

\[ \Delta_c = \frac{1}{\delta_{c}^{M}} \sqrt{(\nabla_x I_a)^2 + (\nabla_y I_a)^2} \]

\[ + \frac{1}{\delta_{c}^{M}} \sqrt{(\nabla_x I_b)^2 + (\nabla_y I_b)^2} \]  

where \( I_a \) and \( I_b \) are the \( a \) and \( b \) channels of \( I_{Lab} \), respectively, and \( \delta_{c,a}^{M} \) and \( \delta_{c,b}^{M} \) are the maximum values of the left and right square root terms, correspondingly. These two maps serve as the next two channels.

3) Disparity information. In our study, we utilize the optical flow method in [50] to estimate the pixel disparity. Here, we use only the displacements in the horizontal axis because we have a strong prior about the disparity. For distorted images, distorted regions may cause severe problems in image fusion such as binocular rivalry or suppression [39], which also affects the attention of humans. Therefore, we refine the estimated disparity information by conducting a correlation measurement between the left and right images.
Fig. 2. Overview of the saliency detection framework. From an S3D pair, cyclopean image $I$ (3 channels), luminance gradient $\Delta_l$, and color gradient $\Delta_c$ are formed. Disparity is also extracted from $I$, and from which normalized disparity $\hat{D}$ (1 channel) and disparity gradient $\Delta_D$ (1 channel) are derived. These form the seven-channel input, which is fed to our deep networks to produce a prediction $\hat{Y}$.

We perform a simple block search for disparity estimation on an S3D pair and if the matched blocks have a low correlation (< 0.4 in our experiment), the disparity value of the block is simply set to zero; otherwise, the value will be one. This strategy is specifically designed to give the deep model some hint about fast-fading distortion (FF), JPEG compression (JPEG) and JPEG2000 compression (JP2K) because these distortions cause many irregularities in images, which eventually affects the estimated disparity. Having the disparity information $D$ of a S3D scene, we define the disparity map as

$$\hat{D} = D_M - D$$

where $D_M = \max_{x,y} D(x,y)$ and $D_m = \min_{x,y} D(x,y)$. Further, the disparity gradient is calculated as

$$\Delta_D = \frac{1}{\delta} \sqrt{(\nabla_x D)^2 + (\nabla_y D)^2}$$

where $\nabla_k D$ denotes the gradient of $D$ along direction $k$ and $\delta$ is the maximum value of the square root term. Ultimately, the seven-channel input can be obtained by concatenating all the above features depth-wisely.

B. Human perception of saliency

To study the behavior of humans viewing images, fixation maps obtained from an eye-tracker, which can locate exactly where humans look in images, are usually used as ground truths. However, these ground truth fixation maps mostly consist of zeros and a few ones in the salient regions. Such sparse data is not proper for most loss functions because it may encourage learning algorithms to optimally produce zeros almost everywhere. Therefore, fixation maps are Gaussian-blurred with $\sigma$ being 1 degree visual angle, which corresponds to the visual acuity in the HVS.

Another problem with the human saliency data is noise. To get rid of noise, one can apply thresholding or perform mean subtraction [3]. However, doing so leads to the risk of losing saliency information induced by distortion, which hurts our purpose. Thus, in this paper, we consider foveation, which refers to the phenomenon that image is sharpest at fovea and blur away as moving to the periphery. Foveation is a process of non-uniform sampling as the photoreceptors are distributed densest at the fovea and quickly decrease in number as one moves from the center of the retina to the periphery. Following [40,41], the foveation model is defined as

$$f(x) = \min \left( \frac{e_2 \ln \left( \frac{1}{\epsilon l_0} \right)}{e (e + e_2)}, \frac{\pi \omega l}{360} \right)$$

where $l$ is the distance between human eye and the fixation point, $CT_0 = 1/64$ is a minimum contrast threshold, $e_2 = 2.3$ is a half-resolution eccentricity constant, $e$ is the eccentricity, and $\epsilon = 0.106$ is a spatial frequency decay constant [12]. The foveation factor can then be defined as

$$F_f(x) = \frac{f(x)}{f(s(x))}$$

where $x$ is a pixel and $s(x)$ is its nearest salient point.

Let us denote $\hat{S}$ the Gaussian-blurred ground truth fixation map. Having defined the foveation, for each saliency map $S$, the final ground truth saliency map is defined as

\[ S_{\text{final}} = \text{f(foveation)}(S) \]
A. Regression convolutional neural network

A region in an image is salient because it possesses not only distinct characteristics in contrast to its local surroundings but also a human-understandable context information. Local information can be accessed via small patches centered at an interested point while larger patch size is required to capture more global and human-interpretable information. The need for a multi-scale neural network arises naturally here. With multiple scales, we expect the model to have a coarse-to-fine look at an image region so that it can capture not only local texture but also more global context information [11].

Based on that reasoning, we propose a three-scale regression CNN for distorted S3D saliency detection (DeepVS-R). The similar use of regression CNN can be traced back to crowd counting [54], but to our knowledge, this is the first work to apply regression to saliency detection. The architecture of our model is shown in Fig. 4. This model consists of two parts: multi-scale and aggregation. The multi-scale part per se is composed of three separate and parallel models. Each model has five convolutional layers and one fully-connected layer.

There is a max pooling operator following the first, second and fifth layers. The activation function is Rectified Linear Unit (ReLU), which is reputable of its biological plausibility [15], for all layers. The outputs of these three models are then concatenated, and this integration comes to the aggregation part. In the aggregation part, there are two fully-connected layers. The first one is an ordinary hidden layer which takes the ReLU as its activation function. The last layer is linear and outputs a single scalar score.

The input data preparation for DeepVS-R is depicted in Fig. 5. After the seven-channel maps are formed, they are resized to 320x640 and split into three-scale patches with a stride of 20 pixels. Each patch centers on each pixel starting from the first pixel of the image. The widths of the first, second and third scales are 50, 100 and 200, respectively. Images are padded with zeros so that patches corresponding to pixels near edges have consistent sizes with the others. In the end, the two larger scales are downsampled to the size of the coarsest one. The saliency score of the pixel at the center of a patch is taken as the value of the dependent variable.

In training, to speed up convergence and avoid overfitting problem, we adopted several techniques frequently used in the deep learning literature. Batch normalization [20] was employed to help gradient flow in back-propagation. As suggested by the authors, the $L_1$ regularization coefficient was set to be small ($10^{-5}$). All the network’s parameters were initialized by Xavier’s method [14]. For the purpose of quickly approaching a good or global minimum, the mean squared error (MSE) loss function was optimized by Adadelta [53]. All the parameters of Adadelta were set to the authors’ suggestion.

B. Fully convolutional neural network

A disadvantage of the regression approach is the requirement of splitting images into patches, which is a significant overhead workload when parallel workers are not available. Thus, we propose to use another model which is a fully convolutional network, which can take a full-resolution feature.
map as input. For this approach, we customized FUCOS in [3] and called it DeepVS-F. The architecture of the model is described in Fig. 6. The model can be viewed as a combination of the convolutional layers from VGG16 network [48] and several transposed-convolutional layers for generating dense prediction. Some customizations include the weights were initialized from those matched of VGG16 [48], and the model was optimized by ADAM [35] rather than stochastic gradient descent. The keeping probability of dropout [49] was set to 0.2. The input of DeepVS-F is the seven-channel maps. In this work, we developed a cost function suitable for saliency detection problems. The cost function is

\[
L(\Theta, \alpha, \beta) = CE(Y, S_{bin}; \Theta) + MSE(\hat{Y}, S; \Theta, \alpha, \beta) + 10^{-4} \times R(\Theta) + 10^{-4} \times s(\hat{Y}, S; \Theta)
\]

(9)

\(S\) is the ground truth saliency map (batch size \(m\), height \(h\), width \(w\)); \(S_{bin}\) is the binary ground truth binarized by a threshold \(\tau\) (we heuristically set to 0.15 in this paper); \(Y\) is the predicted map from DeepVS-F; \(\hat{Y} = \alpha \odot Y(\Theta) + \beta\) (\(\odot\) denotes the Hadamard product) is the parametrized version of \(Y\) and is our final predicted map; \(\alpha\) and \(\beta\) are the parameters of shape \((h, w)\) and to be learned like other network’s parameters; \(\Theta\) is the parameters of the deep network; \(CE(Y, S_{bin}; \Theta)\) is the regular binary cross-entropy; \(MSE(\hat{Y}, S; \Theta, \alpha, \beta)\) is a parametrized MSE cost and is defined as

\[
MSE(\hat{Y}, S; \Theta, \alpha, \beta) = \frac{1}{m \times h \times w} \left\| \hat{Y} - S \right\|_F^2 = \frac{1}{m \times h \times w} \left\| (\alpha \odot Y(\Theta) + \beta) - S \right\|_F^2
\]

(10)

where \(R(\Theta)\) is an \(L_2\) regularization term. The last term is \(s(\hat{Y}, S; \Theta)\) which enforces the gradient similarity and is defined as

\[
s(\hat{Y}, S; \Theta) = \frac{1}{m \times h \times w} \sum_{i=1}^{m} \sum_{d \in \{h,w\}} \left\| \frac{\partial \hat{Y}}{\partial d} - \frac{\partial S}{\partial d} \right\|_F^2
\]

(11)

where \(d\) is the horizontal or vertical direction. The first sum is over all spatial locations and the second one is over \(m\) images in a batch. The final predicted map is \(\hat{Y} = \alpha \odot Y + \beta\). Originally, FUCOS [3] is trained with a MSE cost function but we found that training the DeepVS-F using MSE was unstable. We also considered sigmoid cross-entropy (CE), which is popular in classification tasks. However, we found that this cost function did not give a good performance as well. When the saliency map is binarized, the correlation of the predictions is hurt because the model was not taught to directly predict saliency amplitude. Therefore in this work, we used the proposed combination of both CE and MSE. Instead of a naive sum of the two costs which surely fails because CE and MSE have quite different behaviors in optimization, we parametrized MSE by two trainable masks \(\alpha\) and \(\beta\) whose shapes are the same as input image’s. We analyze the effect of these trainable parameters in Section IV-B.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

We have chosen the IEEE-SA S3D [1] and LIVE1\(^1\) [44] databases. In the former, there are 26 pristine images (ORI) and their five-type distorted versions (Gaussian blur (BLUR), FF, JP2K, JPEG, white Gaussian noise (WN)) at five levels of ascending degree of distortion. In total, there are 676 images in this database. In the latter, there are 20 reference images and 365 distorted images (80 each for JP2K, JPEG, WN and FF; 45 for Blur). Different from the IEEE-SA dataset, LIVE1 does not provide labels for distortion levels. To obtain fixation ground truth data, we conducted an eye-tracking experiment with 20 individuals aged 20 to 30. A “SMART EYE PRO” was used as the binocular eye-tracker combined with a 23” polarization stereoscopic display with resolution of 1600 × 900. The experiment was conducted in a dark room for a fair comparison to the conventional subjective experiments. More information regarding the experiment can be found in [23,26].

To evaluate the models, the training/testing portion was divided as follows: 23/3 for IEEE and 17/3 for LIVE1. The division is carried out with respect to the reference images. To partition the images into training and testing sets, we first randomly divided the pristine images into training and testing portions. Once this was done, all their distorted versions followed them accordingly so that no similar images were in both sets. Since the database is very small, overfitting can
As can be seen from Table II, by our proposed model and the benchmarking algorithms on distortion) and type. Table IV demonstrates the results produced for the seven hand-crafted feature maps based on the analysis of the MIT Saliency benchmark and whose implementations are publicly available. After that, we will demonstrate the need for the seven hand-crafted feature maps based on the analysis of the experiments involving the seven feature maps and other existing deep models. The detailed results are presented below.

A. Comparison with other methods

Fig. 8 shows a visual comparison between our models and some benchmarking algorithms. While other methods struggle to predict the saliency, our two models produce the predictions very close to the ground truth. Tables II and III list our metric performance in the IEEE-SA database according to distortion level (level 0 refers to pristine while level 5 implies high distortion) and type. Table IV demonstrates the results produced by our proposed model and the benchmarking algorithms on both IEEE-SA and LIVE1. As can be seen from Table II, there is a common trend among the bottom-up methods that happen easily. To tackle this problem, for all the weights that matched those of VGG-16, we initialized them from the pretrained VGG-16 on ImageNet. Furthermore, we applied early-stopping in training. Fig. 7 depicts the training and validation curves obtained in our training sessions. As can be seen, the validation loss appears to saturate after epoch 8 and shows an overfitting sign after epoch 12. Therefore, it is reasonable to stop training anytime during this interval. In our implementation, we terminated the training session at epoch 12. Finally, we cross-validated the method five times with different training/testing divisions and averaged the results over all five runs. In practice, the same strategy is frequently occupied in areas having limited data such as image quality assessment [29]–[31,34].

For a qualitative look, we blurred and superimposed the predicted maps over the input images to make heat maps. For quantitative results, we recruited the implementation of the area under the curve by Judd (AUCJ) in [5] and the linear correlation coefficient (CC). Higher scores from these metrics indicate a better performance.

First, we compare our models with four classic methods: GBVS [17], Hou [18], Itti [22] and Fang 3D [10] and five deep learning models: FUCOS [3], SAM-Resnet [8], MLnet [7], Shallow [47] and SaIGAN [46] which are among the leaders of the MIT Saliency benchmark and whose implementations are publicly available. After that, we will demonstrate the need for the seven hand-crafted feature maps based on the analysis of the experiments involving the seven feature maps and other existing deep models. The detailed results are presented below.

\[
D(f_{ref} \| f_{tar}) = E_{f_{ref}}[\log f_{ref}(x) - \log f_{tar}(x)]
\]

where \( x \) is the pixel coordinate, \( f_{ref} \) and \( f_{tar} \) are the reference and target saliency maps, respectively, and \( E_{f_{ref}}[\cdot] \) is the expectation over \( f_{ref} \). In practice, saliency maps are repeatedly subsampled and divided into non-overlapping blocks to perform ToVA in a multi-scale and multi-block fashion. Let \( s \)
denote the number of scales and \( b \) the number of blocks. The unnormalized ToVA score is computed as
The ToVA \(_{\text{unnorm}}(f_{\text{ref}}, f_{\text{tar}}) = \frac{1}{s} \sum_{i,j} D(f_{\text{ref},i,j} \| f_{\text{tar},i,j}) \) (13)

where \(f_{\text{ref},i,j}\) and \(f_{\text{tar},i,j}\) are blocks \(j\) at scale \(j\) from the reference and target maps, respectively. Originally, ToVA is normalized by a constant but we omit it here for the sake of simplicity. It has been shown that the ToVA score is well-correlated with the degradation of images and the visual discomfort level of viewers. Hence, it can play a significant role in image quality/discomfort assessment, image restoration and discomfort reduction.

We show the quantitative results in Figs. 8 and 9. The x-axis indicates distortion level while the y-axis shows the ToVA score. It is clear that the more distorted the images, the more different the predicted saliency maps. The change is significant when the distortion level comes to 4 and 5. This trend is also similar to the one in the previous study [26]. SAM-Resnet produces different saliency maps at different levels but we argue that this is not the ToVA; the model overreacts to the small difference in the input in an unpredictable and unexpected way. Also, there is some change in the saliency maps of other algorithms due to the distorted information they used to predict saliency. Despite that, it is not the ToVA because the maps are not well correlated with the human ground truths in the first place. Because the ToVA metric proportionally reflects the quality and visual discomfort [26], we can interpret our result as an image quality and visual discomfort indicator.

B. Significance of the proposed cost function

To understand how the proposed cost function affects the predicted saliency map, we trained DeepVS-F using several variants of the cost functions. Table V shows the performance...
TABLE V

<table>
<thead>
<tr>
<th>Model</th>
<th>AUC</th>
<th>CC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BLUR FF JP2K JPEG ORI WN Overall</td>
<td>BLUR FF JP2K JPEG ORI WN Overall</td>
</tr>
<tr>
<td>DeepVS-F (CE)</td>
<td>0.77 0.68 0.71 0.78 0.74</td>
<td>0.47 0.41 0.43 0.46 0.46 0.47 0.45</td>
</tr>
<tr>
<td>DeepVS-F (MSE)</td>
<td>0.80 0.70 0.76 0.79 0.82 0.77</td>
<td>0.45 0.45 0.44 0.45 0.43 0.44 0.44</td>
</tr>
<tr>
<td>DeepVS-F (CE+MSE)</td>
<td>0.79 0.72 0.76 0.81 0.80 0.77</td>
<td>0.45 0.46 0.45 0.46 0.44 0.45 0.45</td>
</tr>
<tr>
<td>DeepVS-F (proposed cost)</td>
<td>0.82 0.72 0.78 0.77 0.84 0.78</td>
<td>0.63 0.62 0.64 0.66 0.61 0.63 0.63</td>
</tr>
</tbody>
</table>

In the previous studies, the center-bias prior is heuristically constructed based on an observation that people tend to look at the center of images. By contrast, we do not explicitly model the Center Bias map; it is desired that the network to learn how to correct its own overestimated predictions produced by sigmoid, which more resembles the approach in [37].

C. Significance of the seven-channel feature map

In this section, since the performance of DeepVS-R is only marginally better than that of DeepVS-F but the data making procedure for DeepVS-R is much more time-consuming, we used only DeepVS-F for this analysis. Fig. 10 exhibits the predicted saliency maps of an FF-distorted and a white-noised images, respectively, from DeepVS-F fed by different combinations of the hand-crafted features. When RGB cyclopean is used as input, the network predicts the main object as saliency in Fig. 10(h) but it is different from the human fixation. When only the disparity and disparity gradient information is used, the prediction in Fig. 10(i) seems to be made based on the image depth because most saliency concentrates on the center and pixels having high depth. Under scrutiny, we can see that the network predicts the saliency at locations where the maps are most distorted, i.e. where the values of the maps are zero. When combining with cyclopean images, the results in Fig. 10(j) appear to have more structure information and more correlated to the ground truth. Interestingly, color and luminance gradient maps alone do not really work well in Fig. 10(k), but when the disparity and disparity gradient are additionally considered, the performance becomes much higher as shown in Table VI, and Fig. 10(m), which hints at the dominance of the disparity-related information over the others. On the other hand, using RGB cyclopean with color and luminance information does not bring good performance in Fig. 10(l). A possible reason is that the inputs consist of too much similar distorted information because both the color and luminance are directly derived from the cyclopean images, and the strong correlation in the input actually causes the decrease in performance. Finally, the maps produced when using all seven features shown in Fig. 10(n) can be viewed as a weighted combination of the outputs produced when taking each feature as input. From Table VI and the analysis above, we arrive at the following conclusions:

- **The RGB cyclopean image**: This helps the deep network identify the salient objects and partially the distortion type. However, the prediction is not well-correlated with human data because people do not look at the whole objects but some specific parts only.
- **The color and luminance gradients**: These maps are the main factor to differentiate the distortions. These features
have a competitive quantitative score with the cyclopean but the qualitative result appears to be better. Although these maps increase the overall performance, the gain is not as much as expected.

- **The disparity and disparity gradient:** This information dominates in most cases. As we can see, the combination of these maps and the RGB cyclopean already results in a very good performance. This suggests that the use of RGB-D might be enough in some cases.
- **In conclusion, each feature plays an important role in the performance of the whole model.** When using all the features, the network can compromise the influence of them, which contributes to a better-correlated prediction.

### D. Network dissection

In this section, we want to gain some semantic insight into what deep networks learn with respect to different inputs. Here, we follow [2] to perform a network dissection experiment to study the semantic alignment between the activation maps in DeepVS-F and the ground truth saliency maps. Originally, a network dissection study requires a database containing images labeled at pixel level. Such database does not exist or are not suitable for saliency prediction. However, since the ground truth map conveys salient regions where humans look, these regions must hold some high-level human-interpretable meaning. Hence, this study is able to examine how the network align with high-level concepts in this semantically meaningful sense and further, how well it matches the HVS.

Concretely, for each input image \( I \) from the test set, we gather all the hidden activation maps \( H_c(I) \) of every hidden node \( c \). Next, each map is thresholded by a \( t_c \) so that the top \( p\% \) pixels of each activation map are kept. A lower-resolution map is bilinearly interpolated to match the output resolution before thresholding. We then calculate the interpretability score of each thresholded map \( T_c(I) \) and \( S(I) \) as

\[
IoU_c = \frac{\sum |T_k(I) \cap S(I)|}{\sum |T_k(I) \cup S(I)|}
\]

where \( IoU_c \) is the intersection-over-union score of unit \( c \) and \( |·| \) is the cardinality of a set. If an \( IoU_c \) score exceeds some threshold \( \tau_c \), that activation node is considered to interpret semantic alignment with the corresponding ground truth and we call it an interpretable node. In all the experiments, we set \( p \) to 5 and \( \tau \) 0.04 as suggested in [2].

Having defined the metric, we measure the interpretability of DeepVS-F and FUCOS because they share similar architecture yet different inputs and training schemes. Tables VII and VIII show the interpretability of each layer of the two networks according to distortion type and distortion level. In general, the interpretability of units in DeepVS-F dominates that of units in FUCOS. As stated in [2], the discrimination ability and interpretability, though one does not imply the other, have a positive correlation. By modifying the training settings and using the proposed seven-channel feature maps, we can obtain better discrimination power and hence, better performance. According to Table VII, we can observe that when using only RGB images as input for the deep model, the interpretability of early layers are much less than those of the late layers, which reinforces the observation in [2]. However, this is not
the case when using seven-channel input. In the early and late layers, the interpretability is already high. This score actually decreases in the middle part, which contradicts what was found in [2]. Nevertheless, their study mainly focused on a much shallower model consisting of only five layers.

There are several observations regarding how distortion affects the interpretability. Firstly, in general, BLUR and WN decrease the semantic meaning of the early layers more than the others. Like V1 cortex in the HVS, the early layers extract edges and corners. This information is affected by BLUR and WN more than other distortion types so the interpretability of these layers is lower compared to ORI. FF, JP2K and JPEG usually create more discontinuities in images, which are both discovered by the HVS and the predictive model and hence, the interpretability is higher than the ORI case. Secondly, with regard to distortion level in Table VIII, it is not surprising that more severe distortion leads to lower semantic alignment of the layers. The last observation is that once the interpretability is low in early layers, it will be low through the whole network. This hints at developing good features and learning schemes that better interpretability is constructed in early layers and maintained through the whole network.
V. CONCLUSION

In this study, we have used seven low-level feature maps extracted from luminance, color, and disparity information and effectively integrated them into two deep learning-based models for saliency prediction on distorted S3D images. We successfully utilized the extracted features in a deep learning framework by proposing two models DeepVS-R and DeepVS-F. While DeepVS-R benefits from an explicit multi-scale architecture and bigger database by working with patches, DeepVS-F can achieve a similar performance by fine-tuning a pretrained model but both the preprocessing and training times are significantly shorter. We showed that even though the architectures are straightforward, our model performed better than other methods in the saliency detection problem. We also analyzed how each feature contributes to the overall performance and discovered how semantically each layer of the deep network align with human-understandable concepts.

ACKNOWLEDGMENT

This work was supported by Samsung Research Funding Center of Samsung Electronics under Project Number SRFC-IT11702-08.

REFERENCES

Anh-Duc Nguyen received the B.Eng. degree in automatic control from Hanoi University of Science and Technology, Vietnam in 2015. He is currently pursuing the Ph.D. degree at Yonsei University, South Korea. His research interests include computer vision, image/video analysis, and machine learning.

Heeseok Oh received the B.S., M.S., and Ph.D. degrees in electrical and electronics engineering from Yonsei University, Seoul, South Korea, in 2010, 2012, and 2017, respectively. Currently, he is with Electronics and Telecommunications Research Institute (ETRI), Daejeon, South Korea. His current research interests include 2D/3D image and video processing based on human visual system, computer vision, extended reality and deep generative model.

Haksub Kim received his B.S., M.S. and Ph.D. degrees in electrical and electronic engineering from Yonsei University, Seoul, Korea in 2008, 2011, and 2016, respectively. From 2016 to 2018, he was with Samsung Electronics Company, Ltd., Suwon, Korea, where he was involved in IT & Mobile Communications development. He is currently with Korea Institute of Science and Technology (KIST), Seoul, Korea, where he is working at Center for Imaging Media Research. His research interests include 2D/3D image and video processing based on human visual system, saliency detection based on human visual attention, quality assessment of 2D/3D image and video, deep-learning, cross-layer optimization, and wireless multimedia communications.

Weisi Lin (F’16) received the Ph.D. degree from King’s College London, London, U.K. He is currently an Associate Professor with the School of Computer Engineering, Nanyang Technological University, Singapore. His research interests include image processing, visual quality evaluation, and perception-inspired signal modeling, with over 340 refereed papers published in international journals and conferences.

Dr. Lin is a fellow of Institute of Engineering Technology and an Honorary Fellow of the Singapore Institute of Engineering Technologists. He has been elected as an APSIPA Distinguished Lecturer from 2012 to 2013. He served as a Technical-Program Chair at the Pacific-Rim Conference on Multimedia 2012, the IEEE International Conference on Multimedia and Expo 2013, and the International Workshop on Quality of Multimedia Experience 2014. He has been on the Editorial Board of the IEEE Transactions on Image Processing, the IEEE Transactions on Multimedia from 2011 to 2013, the IEEE Signal Processing Letters, and the Journal of Visual Communication and Image Representation.

Jongyoo Kim received his B.S. degree in Electrical and Electronic Engineering from Yonsei University, Seoul, Korea in 2011, 2013 and 2018, respectively. His research interests include 2D/3D image and video processing based on human visual system, quality assessment of 2D/3D image and video, 3D computer vision and deep learning. He was a recipient of the Global PhD Fellowship by National Research Foundation of Korea from 2011 to 2016. He joined Microsoft Research Asia in March 2018.

Sanghoon Lee (M’05–SM’12) received the B.S. in E.E. from Yonsei University in 1989 and the M.S. and Ph.D. in E.E. from Korea Advanced Institute of Science and Technology (KAIST) in 1991. From 1991 to 1996, he worked for Korea Telecom. He received his Ph.D. in E.E. from the University of Texas at Austin in 2000. From 1999 to 2002, he worked for Lucent Technologies on 3G wireless and multimedia networks. In March 2003, he joined the faculty of the Department of Electrical and Electronics Engineering, Yonsei University, Seoul, Korea, where he is a Full Professor. He was an Associate Editor of the IEEE Trans. Image Processing (2010-2014). He has been an Associate Editor of IEEE Signal Processing Letters (2014-) and Journal of Electronic Imaging (2015-) and Chair of the IEEE P333.1 Quality Assessment Working Group (2011-). He currently serves on the Technical Committee of the IEEE Multimedia Signal Processing (2016-) and the IEEE IVMSP Technical Committee (2014-), was Technical Program Co-chair of the International Conference on Information Networking (ICOIN) 2014, and of the Global 3D Forum 2012 and 2013, and was General Chair of the 2013 IEEE IVMSP Workshop. He also served as a special issue Guest Editor of IEEE Trans. Image Processing in 2013, and as Editor of the Journal of Communications and Networks (JCN) (2009-2015). He received a 2015 Yonsei Academic Award from Yonsei University, a 2012 Special Service Award from the IEEE Broadcast Technology Society and a 2013 Special Service Award from the IEEE Signal Processing Society. His research interests include image/video quality assessment, computer vision, graphics, cloud computing and multimedia communications and wireless networks.