Saliency Preservation in Low-Resolution Grayscale Images

Shivanthan Yohanandan[1], Andy Song[1], Adrian G. Dyer[1], and Dacheng Tao[2]

[1] RMIT University, Melbourne, Australia
{shivanthan.yohanandan,andy.song,adrian.dyer}@rmit.edu.au
[2] UBTECH Sydney AI Centre, SIT, FEIT, The University of Sydney, Australia
dacheng.tao@sydney.edu.au

Abstract. Visual salience detection originated over 500 million years ago and is one of nature's most efficient mechanisms. In contrast, many state-of-the-art computational saliency models are complex and inefficient. Most saliency models process high-resolution color images; however, insights into the evolutionary origins of visual salience detection suggest that achromatic low-resolution vision is essential to its speed and efficiency. Previous studies showed that low-resolution color and high-resolution grayscale images preserve saliency information. However, to our knowledge, no one has investigated whether saliency is preserved in low-resolution grayscale (LG) images. In this study, we explain the biological and computational motivation for LG, and show, through a range of human eye-tracking and computational modeling experiments, that saliency information is preserved in LG images. Moreover, we show that using LG images leads to significant speedups in model training and detection times and conclude by proposing LG images for fast and efficient salience detection.

Keywords: Saliency detection, Fully convolutional network, Peripheral vision

1 Introduction

Visual scenes often contain more items than can be processed concurrently due to the visual system's limited processing capacity [1]. Visual salience (or attention) detection is a cognitive mechanism that efficiently deals with this capacity limitation by selecting relevant or salient information, while ignoring irrelevant information [1]. Salience detection is a fundamental vision mechanism present in many sighted organisms. Even insects, despite having significantly smaller brains and dissimilar eyes to vertebrates, can detect salient stimuli in their visual field [2, 3, 4]. Salience detection can be crudely divided into bottom-up and top-down mechanisms. Bottom-up salience is stimulus and feature-driven, and responsible for automatic, involuntary rapid shifts in attention and gaze. In contrast, top-down salience is task-driven, experience-based, and varies between individuals [5].

Recently, deep neural networks have achieved state-of-the-art performance on various saliency benchmarks [6, 7, 8, 9]. Nevertheless, this success comes at high computational costs [10, 11]. Training and running these networks is time- and resource-intensive, which is not easily scalable to resource-limited devices [10]. Processing high-resolution or stacked multi-resolution color images contributes to these limitations [12].

In contrast, natural visual salience detection proves to be much more efficient. A deeper understanding of the evolutionary origins of visual salience detection suggests that bottom-up saliency is computed from achromatic low-resolution information [13].

Previous studies have shown that low-resolution *color* (LC) [14, 15, 16] and *high-resolution* grayscale (HG) [17, 18, 19, 20, 21, 22] images preserve saliency information, yet are significantly more computationally attractive than high-resolution color (HC) images. Low-resolution grayscale (LG) images are even more computationally attractive, compared to LC and HG images. Nevertheless, to our knowledge, no one has investigated whether saliency information is preserved in LG images. In this study, we therefore investigate saliency preservation in LG images, and present the following three contributions: (1) linking low-resolution grayscale information with the bio-inspired evolutionary origins of visual saliency, (2) assessing the preservation of saliency information in low-resolution grayscale images, and (3) proposing low-resolution grayscale images for fast and efficient saliency detection. Therefore, based on a deeper understanding of the evolutionary origins of visual saliency, together with knowledge gained from studies investigating salience preservation in LC and HG images, we hypothesize that saliency information is well-preserved in LG images.

2 Related Work

2.1 Fixations on Low-Resolution Images

Judd *et al.* [14] investigated how well fixations on LC images predict fixations on the same images in HC. They found that fixations on LC images (76×64 pixels) can predict fixations on HC images (610×512 pixels) quite well (AUC-Judd [14] > 0.85). However, they did not investigate the HC fixation-predictability of LG images, nor did they mention any biological plausibility for deciding to investigate fixations in LC images. Nevertheless, they concluded that working with fixations on LC instead of HC images could be perceptually adequate and computationally attractive, which is part of our motivation for pursuing this study.

2.2 Multi-Resolution Approaches

Deep artificial neural networks are not inherently scale-invariant [23]. Therefore, multi-resolution models are often used to capture saliency at different scales. Advani *et al.* [24] presented a multi-resolution framework for detecting visual salience where resolution degrades further away from the point of fixation represented as a three-level architecture: a central high-resolution fovea (960×960 pixels), a mid-resolution filter (640×640 pixels), and a low-resolution region (480×480 pixels). They found significant computational benefits using this model, but only investigated color images and ignored the achromaticity of peripheral vision.

Shen *et al.* [15] went a step further and modeled the visual acuity of the parafovea and periphery as a stack of multi-scale inputs. They extracted multi-resolution image patches in multiple visual acuity on the same image from fixation targets and non-target locations based on the sunflower model of retina [25, 26, 27]. However, despite finding

comparable performance to higher-resolution models, they too only investigated color images, and overlooked the fact that the parafovea and periphery predominantly processes achromatic information [13]. Furthermore, multi-scale models need to process and train on the same image multiples times at different resolutions, which is computationally unattractive. Therefore, the ideal input image has the lowest resolution and smallest color space that preserves saliency.

2.3 Fixations in Grayscale

Colour processing in chromatic vision conveys processing advantages when combined with brightness information and higher level cognitive influences (e.g. top-down task-driven visual search [28]). Nevertheless, colour information alone is poor at object detection tasks or enabling spatial resolution [29, 30, 31].

Hamel *et al.* [19] investigated the role of color in visual attention by comparing eye movements across different participants viewing color and grayscale videos. They found color to only have a modest effect in predicting salience. However, they only investigated high-resolution images, leaving the influence of color in low-resolution images a gap for us to fill.

Yang *et al.* [32] also investigated whether saliency information is preserved in grayscale images using a novel minimization function. They showed that saliency is well-preserved in grayscale images of the same resolution, but did not extend their investigation to lower resolutions, which our study aims to do.

3 Evolutionary Origin of Visual Saliency

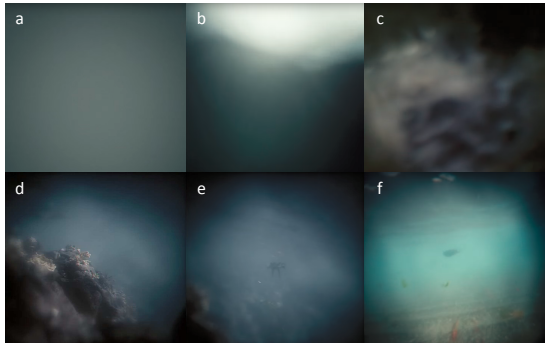

Fig. 1. Hypothetical stages of the evolution of vertebrate vision. This figure panel shows a series of photographic reconstructions of how the vertebrate eye is hypothesized to have evolved, and what that vision hypothetically looked like from an animal's point of view. Static images adapted from *Cosmos: A Spacetime Odyssey (Some of the Things that Molecules do)* [33].

In the beginning, life was blind. Then, around 600 million years ago, the first eyes discriminated night and day (Figure 1(a)) [34]. Light-source localization followed a few million years later (Figure 1(b)), heralding eyes capable of distinguishing light from shadow, thus crudely making-out objects in their vicinity (Figure 1(c)), including those to eat, and those that might eat it. This was likely the birth of stimulus-driven, bottom-up visual salience detection – the mechanism thought to be primarily responsible for the Cambrian explosion [13]. Later, things became a little clearer. The eye's opening contracted to a pinhole covered by a protective transparent membrane, allowing just enough light to paint a dim image on the sensitive inner surface of the eye [35]. Then came focus-sharpening lenses (Figure 1(d)), foveated central vision (Figure 1(e)), and finally, color (Figure 1(f)). However, despite the arrival of high-acuity chromatic central vision, blurry achromatic peripheral vision dominates over 90% of our visual field, and is still the primary information source for bottom-up salience detection – a relic mechanism conserved through evolution in many species because of its apparent speed and efficiency [13]. Furthermore, many sighted animals completely lack chromatic vision, yet are able to rapidly detect obstacles and avoid collisions in complex environments [36].

The ability then of an organism's pupil to rapidly shift foveal gaze to salient regions suggests that it is peripheral vision that points the sharper, high-resolution foveated (sometimes chromatic) vision to investigate objects and regions further. Eye movements align objects with the high-acuity fovea of the retina, making it possible to gather detailed information about the world [35]. Therefore, bottom-up visual salience detection is predominantly a peripheral vision information processing task.

4 Peripheral Vision

A key to the speed and efficiency in bottom-up salience detection lies in the distribution of rod and cone photoreceptor cells in the human retina (Figure 2(a)), and the information processing pipeline of typical vertebrate peripheral vision (Figure 2(b)). Rods primarily encode achromatic luminance (brightness) information, and have a higher distribution outside the fovea. In contrast, cones encode chrominance (color), and are concentrated in the fovea (center of the retina) [37]. Moreover, multiple rods converge to and activate a single retinal ganglion neuron, which is why rod vision has lower spatial resolution compared to information encoded by cones, despite having a high peripheral distribution. In contrast, each cone activates multiple ganglion neurons, resulting in higher acuity vision [39]. Therefore, afferent ganglion neurons, not photoreceptors, from the retina determine the perceived image resolution.

The sparse retinal output of peripheral vision (only 10% of all ganglion cells leaving the eye) enters a structure called the optic tectum (or superior colliculus (SC) in higher-order animals, Figure 2(b)). This structure has only recently emerged as a likely candidate for encoding the saliency map – a well-known precursor for bottom-up salience detection [38, 40, 41]. Furthermore, the SC has direct control of eye muscles. In their study, Veale *et al.* [38] explain that direct retinal input into the SC of a macaque brain can trigger reflex-like saccades via brainstem oculomotor nuclei (red pathway in Figure 2(b)). This could explain why bottom-up saliency detection is rapid and reflex-like,

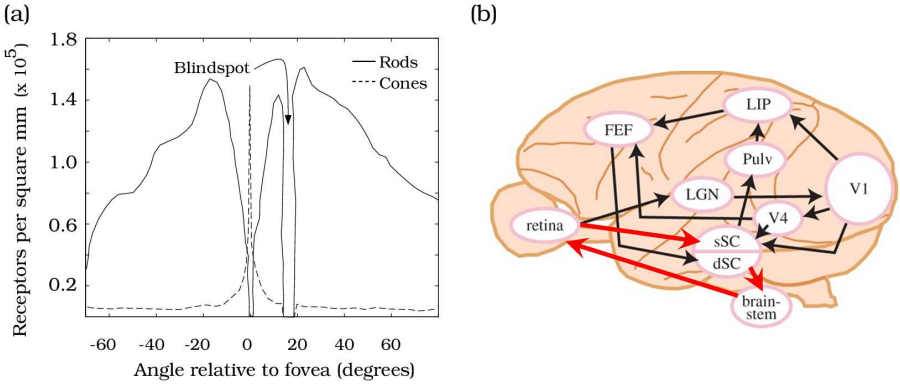

Fig. 2. (a) The human retina's distribution of rod and cone photoreceptors is shown in degrees of visual angle relative to the position of the fovea for the left eye. Cones, concentrated in the fovea, encode high-resolution color. Rod photoreceptors distributed outside the fovea encode low-resolution grayscale information [37]. (b) Macaque brain information flow from retinal input to eye movement output. Visual signals from the retina to the cerebral cortex are mediated through the primary visual cortex (V1) and the superior colliculus (sSC and dSC). There is also a shortcut from the superficial (sSC) to the deep (dSC) superior colliculus, which then sends outputs directly to the brainstem oculomotor nuclei, resulting in rapid saccades (red pathway) [38].

which makes sense since it is processing predominantly achromatic information from fewer afferent neurons, compared to foveated vision, which is processed downstream of the SC and in larger complex brain regions, therefore taking longer. This means far fewer neurons enter the SC, which is analogous to a low-resolution grayscale digital image. Therefore, this sparse achromatic peripheral output could be approximated using low-resolution grayscale images in the digital domain.

5 Approximating Peripheral Vision

Leveraging knowledge from Judd *et al.* [14] and Hamel *et al.* [19], we decided to approximate peripheral vision by first transforming the color space of HC images to 8-bit grayscale (section 5.1), followed by down-sampling the original image height to 64 pixels and width proportionally (section 5.2).

5.1 Colorimetric Grayscale Conversion

Images were first converted from 24-bit sRGB to 8-bit grayscale since it is faster and more efficient to consolidate the three channels before performing subsequent operations, which would otherwise need to be performed thrice (i.e. once per channel). Color to grayscale conversion is a lossy operation, resulting in luminance degradation, which may affect saliency [17]. To avoid such systematic errors, the grayscale conversion must at least preserve the brightness features of the original stimuli (i.e. the luminosity

of grayscale pixels must be identical to the original color image). The HC images used in this study are stored in the sRGB (standard Red Green Blue) color space, which also defines a nonlinear transformation (gamma correction) between the luminosity of these primaries and the actual number stored.

To convert the 24-bit sRGB gamma-compressed color model I_{HC} to an 8-bit grayscale representation of its luminance I_{HG}, the gamma compression function must first be removed via gamma expansion to transform the image to a linear RGB color space [42], so that the appropriate weighted sum can be applied to the linear color components R_{linear}, G_{linear}, B_{linear}. For the sRGB color space, gamma expansion is defined as

$$C_{linear} = \begin{cases} \frac{C_{sRGB}}{12.92} & C_{sRGB} \leq 0.04045 \\ (\frac{C_{sRGB}+0.055}{1.055})^{2.4} & C_{sRGB} > 0.04045 \end{cases} \tag{1}$$

where C_{sRGB} represents any of the three gamma-compressed sRGB primaries (R_{sRGB}, G_{sRGB}, and B_{sRGB}, each in range $[0, 1]$) and C_{linear} is the corresponding linear-intensity value (R_{linear}, G_{linear}, and B_{linear}, also in range $[0, 1]$). Then, I_{HG} is calculated as a weighted sum of the three linear-intensity values, which is given by

$$I_{HG} = 0.2126 \times R_{linear} + 0.7152 \times G_{linear} + 0.0722 \times B_{linear}. \tag{2}$$

These three coefficients represent the intensity (luminance) perception of a standard observer trichromat human to light of the precise Rec. 709 [43] additive primary colors that are used in the definition of sRGB.

5.2 Down-Sampling Image Resolution

We chose 64 pixels as our low-resolution height since Judd *et al.* [14] found this to be the resolution with the best resolution-saliency compromise compared to other resolutions. According to the Nyquist theorem, down-sampling from a higher-resolution image can only be carried out after applying a suitable 2D anti-aliasing filter to prevent aliasing artifacts. To reduce the height of each image down to 64 pixels, we used the same method as Torralba *et al.* [44]: we first applied a low-pass 5×5 binomial filter to I_{HG} and then down-sampled the resulting image using bicubic interpolation by a factor of two, until the desired image height of 64 pixels was reached (corresponding width was maintained based on the original aspect ratio), forming I_{LG}. This also had the effect of providing a clear upper bound on the amount of visual information available [44].

6 Experiments

This section assesses how well saliency information is preserved after transforming HC images to LG images using methods outlined above. Furthermore, it investigates if there are any computational benefits using LG over HC. A fixation map is a two-dimensional spatial record of discrete image locations fixated by an observer, and is collected using an eye-tracker [45]. Previous studies used fixation maps to compare saliency similarity between images [14, 46]. Saliency similarity can also be quantified using fixation-map

inter-observer visual congruency (agreement) [46]. To that end, we designed and conducted three separate experiments: section 6.1 assesses LG and HC fixation-map similarity; section 6.2 assesses LG vs. HC fixation-map inter-observer congruency; and section 6.3 compares accuracy, training and detection speed performance between saliency models trained on HC and LG data.

6.1 HC and LG Fixation-Map Similarity

Dataset. A subset $\boldsymbol{I_{HC}}$ of 20 HC images (1920×1080 pixels, sRGB) along with the corresponding aggregated eye fixations $\boldsymbol{F_{HC}}$ from 18 observers were randomly sampled from the publicly-available CAT2000 benchmark dataset [47]. This dataset contains 4000 images selected from a wide variety of categories such as *art, cartoons, indoor, jumbled, line drawings, random, satellite, and outdoor*. Overall, this dataset contains 20 different categories with 200 images from each category. Only 20 images were evaluated since the sample size of observers was sufficient to determine the statistical significance of fixation-map similarity. Using methods outlined in section 5, images from $\boldsymbol{I_{HC}}$ were first converted to grayscale, then down-sampled to 120×64 pixels. This resulted in a set of images $\boldsymbol{I_{LG}}$ that were a mere 0.12% of the original size, thus significantly reducing computational costs. For human visualization on the eye-tracker screen, $\boldsymbol{I_{LG}}$ images were up-sampled back to their original resolution using the same bicubic interpolation rescaling method outlined in section 5.2.

Eye Tracking. Eye fixations $\boldsymbol{F_{LG}}$ were collected using a Tobii T60 eye-tracker by allowing a separate cohort of 18 consenting participants to free-view each $\boldsymbol{I_{LG}}$ image for 3 seconds from a viewing distance of 60 cm, consistent with the CAT2000 study. Such a viewing duration typically elicits 4-6 fixations from each observer. This is sufficient to highlight a few points of interest per image, and offers a reasonable testing ground for saliency models [48]. Each observer underwent an initial five-point calibration procedure to minimize eye-tracking calibration errors. Every pair of LG/HC images was displayed at least 2 images apart to minimize the effect of priming.

Evaluation Metrics. We compared $\boldsymbol{F_{HC}}$ and $\boldsymbol{F_{LG}}$ fixation map similarity as a function of six recommended "gold standard" metrics: Normalized Scanpath Saliency (NSS) [49], Kullback-Leibler divergence (KL) [50], Judd Area under ROC Curve (jAUC) [51], Shuffled AUC (sAUC) [52], Pearsons Correlation Coefficient (CC) [53], and Similarity or histogram intersection (SIM) [54]. NSS is computed as the average normalized saliency at fixated locations. KL divergence measures the difference between two probability distributions. jAUC measures the area under the Receiver Operating Characteristic curve representing the trade-off between true and false positives at various discrimination thresholds. The sAUC samples negatives from fixation locations from other images, which has the effect of sampling negatives predominantly from the image center. This is because averaging fixations over many images results in the natural emergence of a central Gaussian distribution. Models only predicting the center achieve an sAUC score ¡= 0:5 because at all thresholds they capture as many fixations on the target image as on other images. CC measures statistical Pearsons correlation between two saliency maps. Finally, SIM measures the similarity between two distributions, viewed as histograms. These metrics have been used in the past to evaluate fixation map simi-

Fig. 3. Twenty images from the CAT2000 dataset [46] in high-resolution color (HC) and low-resolution grayscale (LG), and their corresponding fixation maps (from 18 observers each) as a function of σ, where $\sigma \in \{3, 9, 27, 81\}$, from Experiment 1 analyses (section 6.1).

larity because of their easy interpretability [48,52]. We skip explaining these metrics in detail for brevity, and refer readers to the relevant publications.

Discrete fixations from $\boldsymbol{F_{HC}}$ and $\boldsymbol{F_{LG}}$ are converted into continuous distribution maps $\boldsymbol{M_{HC}}$ and $\boldsymbol{M_{LG}}$, respectively, by smoothing, which acts as regularization, allowing for uncertainty in the ground truth measurements to be incorporated. A blur value σ is required for the Gaussian low-pass filter in the Fourier domain. We follow common practice [48], and blur each fixation location using a Gaussian with σ ranging from 1 to 100, resulting in 100 fixation maps for each HC and LG image per participant. For highly similar fixation maps, all evaluation metrics rise (except KL, which falls) rapidly towards a large maximum as $\sigma \rightarrow 100$. Conversely, for highly dissimilar fixation maps, evaluation metrics decrease with an increasing σ [55]. We calculated these metrics using MATLAB scripts from [48], and plot the median across all participants for each metric (Figure 4).

Results. From visual inspection (Figure 3), we can see that increasing σ smooths the fixation density map and has the effect of filtering out stray fixations with low inter-observer congruency, leaving behind high-confidence fixations. These results suggest that $\boldsymbol{M_{HC}}$ and $\boldsymbol{M_{LG}}$ are highly similar, attaining high jAUC (0.88), SIM (0.85) and CC (0.92) as $\sigma \rightarrow 100$ (Figure 4). Moreover, this result confirms saliency preservation in LG images in terms of fixation map similarity.

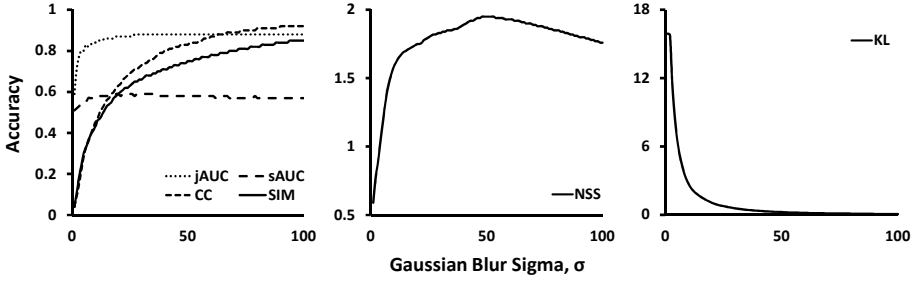

Fig. 4. Low-resolution grayscale and high-resolution color fixation-map similarity metrics as a function of the Gaussian blur σ. Plots represent medians across all participants for all 20 images. Note: NSS y–axis range is constrained to min/max and all plots share the x–axis.

6.2 HC vs. LG Inter-Observer Consistency

Dataset. To determine LG and HC inter-observer congruency (agreement), a subset $\boldsymbol{I_{HC}}$ of 10 HC (1280×1024 pixels, sRGB) images were randomly sampled from the Internet (Google Images) and converted to 120×64 pixel LG images $\boldsymbol{I_{LG}}$ using the same methods described in section 5. As with the previous experiment (section 6.1), 10 images were deemed sufficient to determine statistical significance since the sample size of observers was large. This resulted in images that were only 0.19% of the original size; once again, significantly reducing computational costs.

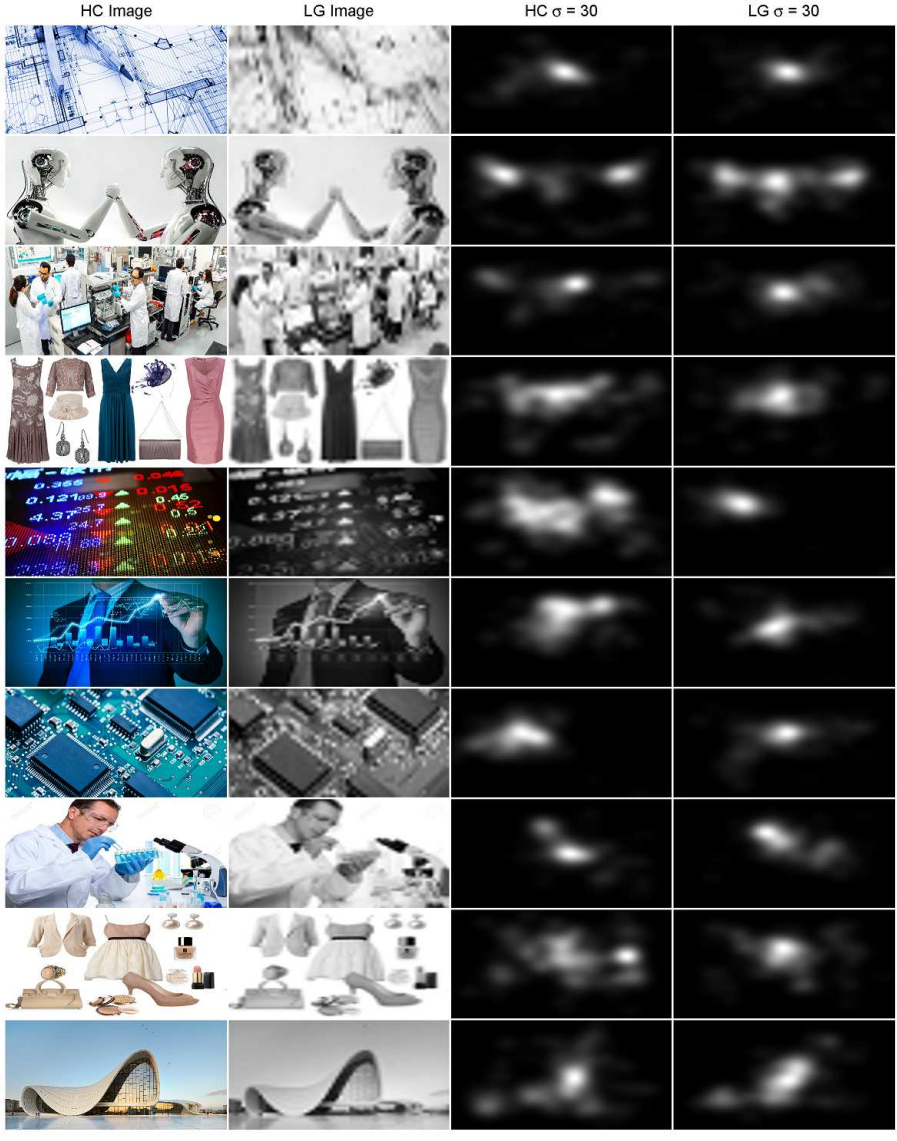

Fig. 5. Full set of 10 images in high-resolution color (HC) and low-resolution grayscale (LG), and their corresponding fixation maps (from 35 observers each) as a function of $\sigma = 30$, from Experiment 2 analyses (section 6.2).

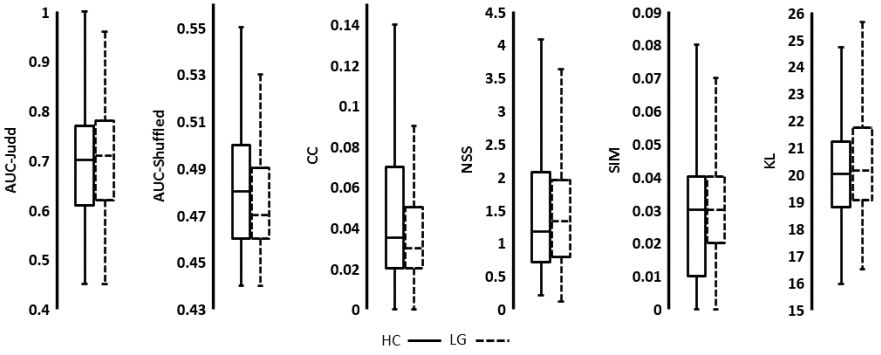

Fig. 6. Experiment 2 (section 6.2) boxplots showing high-resolution color (HC) vs. low-resolution grayscale (LG) inter-observer congruency across 35 observers across 10 images (in HC and LG) as a function of AUC-Judd, AUC-Shuffled, CC, NSS, SIM, and KL. A σ of 30 was chosen from Experiment 1 to generate the fixation maps used in this analysis. ANOVA analysis revealed no statistically significant difference between HC and LG across all 6 metrics. Note: y–axes have been cropped and scaled for viewing convenience; boxplot key (from bottom): minimum, 25^{th} percentile, median, 75^{th} percentile, and maximum.

Eye Tracking. To conduct this analysis, we required separate fixation data from each observer, which was lacking from the CAT2000 dataset's aggregated fixations. To that end, we collected eye-tracking fixation data $\boldsymbol{F_{HC}}$ and $\boldsymbol{F_{LG}}$ using the same Tobii eye-tracker from 35 consenting observers viewing both sets of $\boldsymbol{I_{HC}}$ and $\boldsymbol{I_{LG}}$ images, respectively. Standard five-point eye-tracker calibration was performed at the start of each trial for each participant as standard practice. Similar to the previous experiment in 6.1, images were presented for 3 seconds each, and participants were instructed to freely view images, while seated 60 cm in front of the screen.

Evaluation Metrics. We chose a Gaussian blur σ of 30, which corresponds to 1 degree of visual angle [48], generated continuous fixation maps, $\boldsymbol{M_{HC}}$ and $\boldsymbol{M_{LG}}$, and calculated inter-observer congruency as a function of the same previous set of 6 metrics within the $\boldsymbol{F_{HC}}$ and $\boldsymbol{F_{LG}}$ sets using the leave-one-out (one-vs-all) method described in [46]. We also performed an ANOVA analysis across all co-variates.

Results. Figure 6 show that the LG fixation data does not show a higher dispersion between observers' eye tracking data compared to HG fixations. Furthermore, the ANOVA analysis found no significant difference between HC and LG inter-observer consistency ($p > 0.05$). This result suggests that LG fixation data is as accurate as expected for substituting HC fixation data [46]. Moreover, this result further confirms saliency preservation in LG images in terms of fixation map inter-observer congruency.

6.3 HC vs. LG Saliency Detection Models

Model Architecture. Conventional convolutional neural networks (CNNs) used for image classification consists of convolutional layers followed by fully connected layers,

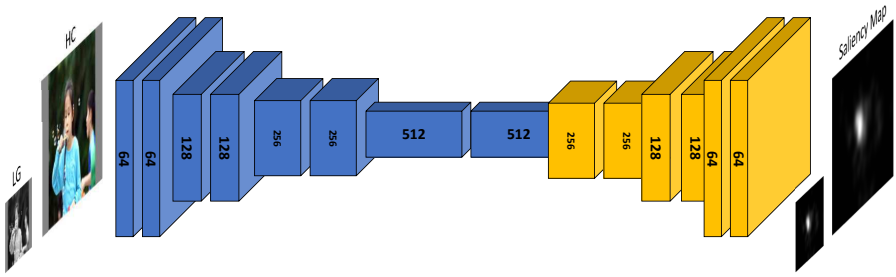

Fig. 7. Fully convolutional neural network architecture. The network takes an HC or LG image as input, adopts convolution layers (blue) with 3×3 kernels and a stride of 1 to transform the image into multidimensional feature representations, then applies a stack of deconvolution layers (orange) for upsampling the extracted coarse features. Finally, a fully convolution layer with a 1×1 kernel and sigmoid activation function outputs a pixel-wise probability (saliency) map the same size as the input, where larger values correspond to higher saliency. Numbers represent convolutional filters.

which takes an image of fixed spatial size as input and produces a single-dimensional vector indicating the class-probability or category of the input image. For tasks requiring spatial labels, like generating pixel-wise saliency heatmaps, we consider fully convolutional neural networks (FCNs) with deconvolutional layers. This architecture has been previously used for saliency detection in video with enormous success [56], which is why we used a slightly modified version in our study (Figure 7). It is capable of generating saliency maps the same size as the input image, which was ideal for our experiment since we needed to compare the same model on datasets comprising images of different resolutions and color spaces without needing to change model hyperparameters. To test if HC and LG models have similar accuracy, we kept all other parameters constant and only varied the image resolution and color space during compression. We were only interested in a HC saliency detection model with comparable accuracy and performance to the state-of-the-art so we could show that an LG model can achieve the same performance faster and more efficiently. The model generates a saliency heatmap from a given input, which can then be compared with the ground-truth density map, just as in the above experiments.

Dataset. The 2000 labeled images from the same CAT2000 saliency benchmark dataset used previously was split into training (1800 images) and validation (200 images) sets. These sets were duplicated and preprocessed to produce four new sets: high-resolution 24-bit color training and validation sets, $\boldsymbol{T_{HC}}$ and $\boldsymbol{V_{HC}}$, created by downsampling the original resolution to 512×512 pixels (typical resolution used by many state-of-the-art saliency detection models) using methods described in section 5.2, and low-resolution (64×64 pixels) 8-bit grayscale sets, $\boldsymbol{T_{LG}}$ and $\boldsymbol{V_{LG}}$, generated using methods outlined above.

Model Training. The Python Keras API with the TensorFlow framework backend was used to implement and train HC and LG FCN models, $\boldsymbol{M_{HC}}$ and $\boldsymbol{M_{LG}}$,

on the respective 1800 training images end-to-end and from scratch (i.e. randomized initial weights). Network weights and parameters were initialized by seeding a pseudo-random number generator with the same seed for all training sessions and models to ensure everything else remained constant. The training images were propagated through the FCN in batches of 8 and 64 for $\boldsymbol{M_{HC}}$ and $\boldsymbol{M_{LG}}$, respectively. Due to the FCN's large parameter space, $\boldsymbol{M_{HC}}$ batch size was restricted to 8 so that the 512×512 images could be accommodated by the available memory (12 GB) and resources. Weights were learned using slow gradient decent (RMSProp) over 100 epochs totaling 180,000 iterations. The base learning rate was set to 0.05, and decreased by a factor of 10 after 2000 iterations. A mean-squared error loss function was implemented to compute loss for gradient descent. An NVIDIA Tesla K80 GPU was used for training and inference. Training time (i.e. the time taken to complete all iterations to completion) for each model was recorded.

Evaluation Metrics. $\boldsymbol{M_{HC}}$ and $\boldsymbol{M_{LG}}$ were tested on their respective held-out validation sets, $\boldsymbol{V_{HC}}$ and $\boldsymbol{V_{LG}}$. The predicted labels from the models' output were upsampled to match the original dimensions of the ground truth labels (1920×1080 pixels) for a fair accuracy evaluation. Model accuracy was defined as a function of NSS, Judd-AUC, SIM, and CC, described above, and computed using MATLAB code from the MIT saliency benchmark GitHub repository [14]. Furthermore, detection time, defined as the average time taken by the model to generate a predicted saliency map based on each of the 200 test images, was also measured for $\boldsymbol{M_{HC}}$ and $\boldsymbol{M_{LG}}$. Two-tailed paired Students *t*-tests were performed between HC and LG result pairs to determine if differences were statistically significant. Finally, to rule out centre bias [57] we applied a 2D Gaussian located at the image centre and statistically compared (using paired Students t-tests) its sAUC with our LG and HC models on 200 test images.

Results. Figures 8(a) and 8(b) show no statistically significant difference between $\boldsymbol{M_{HC}}$ and $\boldsymbol{M_{LG}}$ accuracy across all evaluation metrics ($p > 0.05$). Furthermore, these accuracies are comparable to state-of-the-art models. The centre-bias sAUC results (2D Gaussian HC = 0.45 and LG = 0.44; our FCN HC = 0.58 and LG = 0.57; p-value < 0.05) discard the hypothesis that our models only predict central saliency. Moreover, an sAUC of 0.58 is highly comparable to state-of-the-art models [58]. Therefore, this is further evidence suggesting saliency is well-preserved in LG images. Figures 8(c) and 8(d) show a significant difference between $\boldsymbol{M_{HC}}$ and $\boldsymbol{M_{LG}}$ training and detection times ($p < 0.05$). $\boldsymbol{M_{LG}}$ trained more than $14\times$ faster than its HC counterpart, $\boldsymbol{M_{HC}}$. Furthermore, $\boldsymbol{M_{LG}}$ is capable of generating a predicted saliency map almost $10\times$ faster than $\boldsymbol{M_{HC}}$ (12 vs. 114 milliseconds). Considering these significant speedups come at negligible accuracy cost, the implications of using LG images over HC are substantial; thus, the motivation to use LG images in saliency detection should now be more obvious and appealing.

7 Conclusion

In this study, we explained and demonstrated the biological and computational motivation for using LG images in salience detection. We learned, through evolutionary insights, that bottom-up visual salience detection is predominantly a peripheral vision

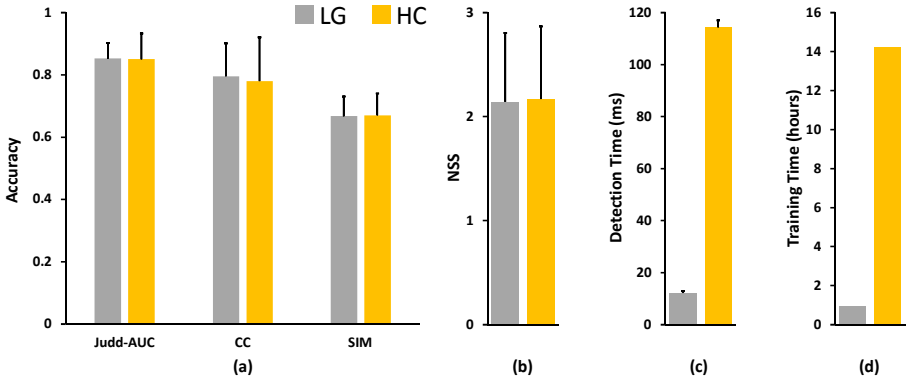

Fig. 8. (a) HC and LG model accuracy as a function of Judd-AUC, CC, SIM, and (b) NSS. (c) HC vs. LG training time. (d) HC vs. LG detection time. Bar plots represent means and error bars represent standard deviations across the 200 test results per model.

mechanism. We also learned that peripheral vision information is primarily achromatic and low-resolution, and can be approximated in the digital domain using a simple LG transformation. Through eye-tracking experiments, we found high similarity between LG and HC fixations. The results of this study also showed no significant difference in inter-observer congruency between LG and HC groups. Additionally, we trained fully convolutional neural networks for saliency detection using LG and HC data from a benchmark dataset and found no significant difference between HC, LG and state-of-the-art model accuracy. However, we found that the LG model required significantly less $(1/14)$ training time and is much faster (almost $10\times$) performing detection compared to the same network trained and evaluated on HC images. Therefore, these results confirm our hypothesis that saliency information is preserved in LG images, and we conclude by proposing LG images for fast and efficient saliency detection. Future research will extend this work by investigating the use of LG images in other computer vision tasks, such as object detection, pose tracking and background subtraction, since we have reason to believe that many vision tasks could just as easily be done using peripheral vision and hence, low-resolution grayscale information.

8 Acknowledgements

This research was supported by an Australian Postgraduate Award scholarship, the Professor Robert and Josephine Shanks scholarship, and Australian Research Council grants FL-170100117, DP-180103424, and LP-150100671. The authors wish to thank the eye tracking participants for volunteering their time and Mr Wei Li for helping with the experiments.

References

1. McMains, S.A., Kastner, S.: Visual Attention. In Binder, M.D., Hirokawa, N., Windhorst, U., eds.: Encyclopedia of Neuroscience. Springer Berlin Heidelberg (2009) 4296–4302
2. Morawetz, L., Spaethe, J.: Visual attention in a complex search task differs between honeybees and bumblebees. The Journal of Experimental Biology **215**(Pt 14) (July 2012) 2515–2523
3. Avargus-Weber, A., Dyer, A.G., Ferrah, N., Giurfa, M.: The forest or the trees: preference for global over local image processing is reversed by prior experience in honeybees. Proceedings of the Royal Society B: Biological Sciences **282**(1799) (January 2015)
4. Morawetz, L., Svoboda, A., Spaethe, J., Dyer, A.G.: Blue colour preference in honeybees distracts visual attention for learning closed shapes. Journal of Comparative Physiology. A, Neuroethology, Sensory, Neural, and Behavioral Physiology **199**(10) (October 2013) 817–827
5. Hou, W., Gao, X., Tao, D., Li, X.: Visual saliency detection using information divergence. Pattern Recognition **46**(10) (October 2013) 2658–2669
6. Kmmerer, M., Wallis, T.S.A., Bethge, M.: DeepGaze II: Reading fixations from deep features trained on object recognition. arXiv:1610.01563 [cs, q-bio, stat] (October 2016)
7. Huang, X., Shen, C., Boix, X., Zhao, Q.: SALICON: Reducing the Semantic Gap in Saliency Prediction by Adapting Deep Neural Networks. In: 2015 IEEE International Conference on Computer Vision (ICCV). (December 2015) 262–270
8. Kruthiventi, S.S.S., Ayush, K., Babu, R.V.: DeepFix: A Fully Convolutional Neural Network for Predicting Human Eye Fixations. IEEE Transactions on Image Processing **26**(9) (September 2017) 4446–4456
9. Cornia, M., Baraldi, L., Serra, G., Cucchiara, R.: Predicting Human Eye Fixations via an LSTM-based Saliency Attentive Model. arXiv:1611.09571 [cs] (November 2016)
10. Rajankar, O.S., D.Kolekar, U.: International Journal of Image, Graphics and Signal Processing(IJIGSP). International Journal of Image, Graphics and Signal Processing(IJIGSP) **7**(8) 58
11. Wang, W., Shen, J.: Deep Visual Attention Prediction. arXiv:1705.02544 [cs] (May 2017)
12. Vo, A.V., Truong-Hong, L., Laefer, D.F., Tiede, D., dOleire Oltmanns, S., Baraldi, A., Shimoni, M., Moser, G., Tuia, D.: Processing of Extremely High Resolution LiDAR and RGB Data. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **9**(12) (December 2016) 5560–5575
13. Lamb, T.D.: Evolution of Phototransduction, Vertebrate Photoreceptors and Retina. In Kolb, H., Fernandez, E., Nelson, R., eds.: Webvision: The Organization of the Retina and Visual System. University of Utah Health Sciences Center, Salt Lake City (UT) (1995)
14. Judd, T., Durand, F., Torralba, A.: Fixations on low-resolution images. Journal of Vision **11**(4) (April 2011) 1–20
15. Shen, C., Huang, X., Zhao, Q.: Learning of Proto-object Representations via Fixations on Low Resolution. ArXiv e-prints **1412** (December 2014) arXiv:1412.7242
16. Ho-Phuoc, T., Guyader, N., Landragin, F., Gurin-Dugu, A.: When viewing natural scenes, do abnormal colors impact on spatial or temporal parameters of eye movements? Journal of Vision **12**(2) (February 2012) 4–4
17. Hamel, S., Guyader, N., Pellerin, D., Houzet, D.: Contribution of Color Information in Visual Saliency Model for Videos. In: Image and Signal Processing. Lecture Notes in Computer Science, Springer, Cham (June 2014) 213–221
18. Hamel, S., Guyader, N., Pellerin, D., Houzet, D.: Contribution of color in saliency model for videos. Signal, Image and Video Processing **10**(3) (March 2016) 423–429

19. Hamel, S., Houzet, D., Pellerin, D., Guyader, N.: Does color influence eye movements while exploring videos? Journal of Eye Movement Research **8**(1) (April 2015)
20. Frey, H.P., Honey, C., Knig, P.: What's color got to do with it? The influence of color on visual attention in different categories. Journal of Vision **8**(14) (October 2008) 6–6
21. Baddeley, R.J., Tatler, B.W.: High frequency edges (but not contrast) predict where we fixate: A Bayesian system identification analysis. Vision Research **46**(18) (September 2006) 2824–2833
22. Dorr, M., Vig, E., Barth, E.: Colour Saliency on Video. In: Bio-Inspired Models of Network, Information, and Computing Systems. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, Springer, Berlin, Heidelberg (December 2010) 601–606
23. Xu, Y., Xiao, T., Zhang, J., Yang, K., Zhang, Z.: Scale-Invariant Convolutional Neural Networks. arXiv:1411.6369 [cs] (November 2014)
24. Advani, S., Sustersic, J., Irick, K., Narayanan, V.: A multi-resolution saliency framework to drive foveation. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. (May 2013) 2596–2600
25. Lindeberg, T., Florack, L.: Foveal scale-space and the linear increase of receptive field size as a function of eccentricity. (1994)
26. Koenderink, J.J., Doorn, A.J.v.: Visual detection of spatial contrast; Influence of location in the visual field, target extent and illuminance level. Biological Cybernetics **30**(3) (September 1978) 157–167
27. Romeny, B.M.H.: Front-End Vision and Multi-Scale Image Analysis: Multi-scale Computer Vision Theory and Applications, written in Mathematica. Springer Science & Business Media (October 2008)
28. Treisman, A.M., Gelade, G.: A feature-integration theory of attention. **12**(1) 97–136
29. Humphrey, G.K., Goodale, M.A., Jakobson, L.S., Servos, P.: The role of surface information in object recognition: studies of a visual form agnosic and normal subjects. **23**(12) 1457–1481
30. Gegenfurtner, K.R., Rieger, J.: Sensory and cognitive contributions of color to the recognition of natural scenes. **10**(13) 805–808
31. Lennie, P.: Color vision: Putting it together. **10**(16) R589–R591
32. Yang, Y., Song, M., Bu, J., Chen, C., Jin, C.: Color to Gray: Attention Preservation. In: 2010 Fourth Pacific-Rim Symposium on Image and Video Technology. (November 2010) 337–342
33. Pope, B., Druyan, A., Soter, S., deGrasse Tyson, N., Hanich, L., Holtzman, S.: Cosmos: A Spacetime Odyssey (episode 2: some of the things that molecules do) (2014)
34. Nilsson, D.E.: The evolution of eyes and visually guided behaviour. Philosophical Transactions of the Royal Society of London B: Biological Sciences **364**(1531) (October 2009) 2833–2847
35. Potter, M.C., Wyble, B., Hagmann, C.E., McCourt, E.S.: Detecting meaning in RSVP at 13 ms per picture. Attention, Perception, & Psychophysics **76**(2) (February 2014) 270–279
36. Stojcev, M., Radtke, N., D'Amaro, D., Dyer, A.G., Neumeyer, C.: General principles in motion vision: color blindness of object motion depends on pattern velocity in honeybee and goldfish. Visual Neuroscience **28**(4) (July 2011) 361–370
37. Wandell, B.A.: Foundations of Vision. Sinauer Associates (January 1995)
38. Veale, R., Hafed, Z.M., Yoshida, M.: How is visual salience computed in the brain? Insights from behaviour, neurobiology and modelling. Philosophical Transactions of the Royal Society B: Biological Sciences **372**(1714) (February 2017)
39. Okawa, H., Sampath, A.P.: Optimization of single-photon response transmission at the rod-to-rod bipolar synapse. Physiology (Bethesda, Md.) **22** (August 2007) 279–286

40. White, B.J., Kan, J.Y., Levy, R., Itti, L., Munoz, D.P.: Superior colliculus encodes visual saliency before the primary visual cortex. Proceedings of the National Academy of Sciences **114**(35) (August 2017) 9451–9456
41. Krauzlis, R.J., Lovejoy, L.P., Znon, A.: Superior Colliculus and Visual Spatial Attention. Annual Review of Neuroscience **36**(1) (2013) 165–182
42. Poynton, C.A.: Rehabilitation of gamma. International Society for Optics and Photonics **3299** (July 1998) 232–250
43. Poynton, C., Funt, B.: Perceptual uniformity in digital image representation and display. Color Research & Application **39**(1) (February 2014) 6–15
44. Torralba, A.: How many pixels make an image? Visual Neuroscience **26**(1) (February 2009) 123–131
45. Wooding, D.S.: Fixation Maps: Quantifying Eye-movement Traces. Proceedings of the 2002 Symposium on Eye Tracking Research & Applications (2002) 31–36
46. Tavakoli, H.R., Ahmed, F., Borji, A., Laaksonen, J.: Saliency Revisited: Analysis of Mouse Movements versus Fixations. arXiv:1705.10546 [cs] (May 2017)
47. Borji, A., Itti, L.: CAT2000: A Large Scale Fixation Dataset for Boosting Saliency Research. arXiv:1505.03581 [cs] (May 2015)
48. Bylinskii, Z., Judd, T., Oliva, A., Torralba, A., Durand, F.: What do different evaluation metrics tell us about saliency models? arXiv:1604.03605 [cs] (April 2016)
49. Peters, R.J., Iyer, A., Itti, L., Koch, C.: Components of bottom-up gaze allocation in natural images. Vision Research **45**(18) (August 2005) 2397–2416
50. Liang, J., Zhang, Y.: Top down saliency detection via Kullback-Leibler divergence for object recognition. In: 2015 International Symposium on Bioelectronics and Bioinformatics (ISBB). (October 2015) 200–203
51. Judd, T., Ehinger, K., Durand, F., Torralba, A.: Learning to predict where humans look. In: 2009 IEEE 12th International Conference on Computer Vision. (September 2009) 2106–2113
52. Borji, A., Tavakoli, H.R., Sihite, D.N., Itti, L.: Analysis of Scores, Datasets, and Models in Visual Saliency Prediction. In: 2013 IEEE International Conference on Computer Vision. (December 2013) 921–928
53. Le Meur, O., Le Callet, P., Barba, D.: Predicting visual fixations on video based on low-level visual features. Vision Research **47**(19) (September 2007) 2483–2498
54. Koch, C., Ullman, S.: Shifts in selective visual attention: towards the underlying neural circuitry. Human Neurobiology **4**(4) (1985) 219–227
55. Engelke, U., Liu, H., Wang, J., Le Callet, P., Heynderickx, I., Zepernick, H.J., Maeder, A.: A Comparative Study of Fixation Density Maps. IEEE Transactions on Image Processing **22**(3) (March 2013) pp.1121–1133
56. Wang, W., Shen, J., Shao, L.: Video salient object detection via fully convolutional networks. **27**(1) 38–49
57. Tatler, B.W.: The central fixation bias in scene viewing: selecting an optimal viewing position independently of motor biases and image feature distributions. **7**(14) 4.1–17
58. Judd, T., Durand, F., Torralba, A.: A benchmark of computational models of saliency to predict human fixations