SAND2017-8681J

# Data Visualization Saliency Model: A Tool for Evaluating Abstract Data Visualizations

Laura E. Matzen, Michael J. Haass, Kristin M. Divis, Zhiyuan Wang, and Andrew T. Wilson

**Abstract**—Evaluating the effectiveness of data visualizations is a challenging undertaking and often relies on one-off studies that test a visualization in the context of one specific task. Researchers across the fields of data science, visualization, and human-computer interaction are calling for foundational tools and principles that could be applied to assessing the effectiveness of data visualizations in a more rapid and generalizable manner. One possibility for such a tool is a model of visual saliency for data visualizations. Visual saliency models are typically based on the properties of the human visual cortex and predict which areas of a scene have visual features (e.g. color, luminance, edges) that are likely to draw a viewer's attention. While these models can accurately predict where viewers will look in a natural scene, they typically do not perform well for abstract data visualizations. In this paper, we discuss the reasons for the poor performance of existing saliency models when applied to data visualizations. We introduce the Data Visualization Saliency (DVS) model, a saliency model tailored to address some of these weaknesses, and we test the performance of the DVS model and existing saliency models by comparing the saliency maps produced by the models to eye tracking data obtained from human viewers. Finally, we describe how modified saliency models could be used as general tools for assessing the effectiveness of visualizations, including the strengths and weaknesses of this approach.

**Index Terms**—Visual saliency, evaluation, eye tracking

## Introduction

Vision is the dominant sense for humans [2], with researchers estimating that over 50% of the brain is involved in processing visual information [1,39]. Given how heavily most humans rely on vision to navigate and understand the physical world, it is no surprise that visualizations are a common tool for helping people to navigate through information. Visualizations leverage the capabilities of the human visual system and can provide users with a natural way to explore and comprehend large amounts of information. However, visualizations can also be confusing and misleading, particularly for complex, multidimensional data sets that do not have a natural visual representation.

Evaluating the effectiveness of visualizations can be very challenging [10,30]. Ideally, visualizations would be evaluated with well-designed user studies, but these are not always possible (e.g. if the designer does not have access to the end users) and can also be expensive and time consuming. It would be useful for designers to have more evaluation tools that can be deployed rapidly and iteratively during the design process to assess visualizations prior to conducting a user study. Prior work has suggested that visual saliency models could be one such tool [26,38].

Visual saliency models assess the visual features of an image to predict which areas of that image will draw a viewer's attention. Saliency models are typically inspired by the structure and function of the human visual cortex. The models take an input image and generate a saliency map that predicts which regions of the image will be most likely to draw a human viewer's attention [24]. There are a variety of metrics that can be used to assess the performance of the models by comparing the saliency maps to human fixation data recorded via eye tracking [4,7,8]. Saliency models have been the subject of a great deal of research in the fields of cognitive science and computer vision, and they could prove useful to visualization designers as well. Since data visualizations make use of the human visual system to convey information, evaluation techniques that are rooted in neural processes could provide useful, generalizable metrics.

It is important to note that saliency models' predictions of where viewers will look are based only on the physical properties of the visual stimulus. They are models of what is known as *bottom-up* visual attention. In real-world tasks, a viewer's eye movements are also guided by *top-down* visual attention, which is influenced by the viewer's goals, expectations, and experience [12,43,46]. In the brain, these two processes operate in parallel. Bottom-up visual attention is drawn to regions of a stimulus that are distinct from things around them in terms of their basic visual features (e.g. contrast, color, motion), and top-down visual attention is allocated voluntarily based on the viewer's task and prior knowledge. Regions with high bottom-up saliency may or may not be relevant to the viewer's task and goals, so there is a constant interplay between the two neural systems that guide visual attention and eye movements [41].

When a saliency model is applied to an image, it produces a map that predicts which regions of the image are most likely to draw the viewer's bottom-up attention. In the context of data visualizations, this could allow designers to assess whether or not their design will draw attention to the most important information [26]. In other words, saliency maps provide designers with a metric of how well bottom-up attention and top-down goals will overlap for the application that the designer has in mind. From the perspective of a person using a visualization, a strong overlap between visual saliency and important features will allow the user to complete tasks faster and more efficiently, minimizing distraction from unimportant information.

Although generating saliency maps for data visualizations could provide a useful and widely applicable evaluation metric, there is a substantial obstacle to this approach. The existing models of bottom-up visual saliency were designed for images of natural scenes, and the visual and spatial properties of natural scenes can be quite different from those of visualizations. While saliency models can generate reasonable predictions of where people will look in scene-like visualizations (i.e., visualizations that resemble photographs) [38], these models typically underperform for abstract visualizations [18].

This is a disadvantage for existing saliency models, but it raises the possibility that these models can be modified to better account for patterns of attention in data visualizations. The differing nature of


- *Laura Matzen, Sandia National Laboratories. E-mail: lematze@sandia.gov.*
- *Michael Haass, Sandia National Laboratories. E-mail: mjhaass@sandia.gov.*
- *Kristin Divis, Sandia National Laboratories. E-mail: kmdivis@sandia.gov.*
- *Zhiyuan Wang, University of Illinois at Urbana-Champaign. E-mail: zhiyuanwang42@gmail.com.*
- *Andrew Wilson, Sandia National Laboratories. E-mail: atwilso@sandia.gov*


Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org.
Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx/.

visualizations and natural scenes also presents opportunities to incorporate some information about top-down attention into saliency models. In the context of natural scenes, top-down attention is highly task- and situation- dependent, making it very difficult to model in any generalized way. This is the reason that most existing saliency models take only bottom-up attention into account. However, in the context of data visualizations, the visual features and their placement within the scene are selected by a designer in support of a particular goal or goals. A designer is structuring the image in order to convey information, so the visual features that the designer selects encode top-down information in a way that the features of a natural scene do not. Visualizations are also typically "born digital," unlike images of natural scenes, making it easier to isolate distinct elements (such as individual data regions or text regions) and infer their importance from a top-down perspective.

In this paper, we explore why existing saliency models underperform for abstract data visualizations. We identify the visual and structural features of visualizations that are incompatible with the existing, scene-based visual saliency models. We then discuss the development of a modified saliency model that addresses these features and incorporates new information based on top-down attention, allowing it to make more accurate predictions of which regions of a visualization will draw a viewer's attention. We outline the features of the Data Visualization Saliency (DVS) model and compare its performance to a set of existing saliency models. Finally, we discuss how the DVS model could be used as an evaluation tool during the process of designing a visualization, allowing designers to rapidly assess how various design choices affect the saliency of different parts of a visualization.

## 1 Evaluation of Existing Saliency Models

There are numerous bottom-up saliency models that have been developed to predict where people will look in natural scenes. Many of these models are based on the neurophysiology of human and other primates' visual systems [3]. They select visual features that are known to elicit neural responses in the visual cortex, such as luminance, hue, contrast and orientation. The feature maps are often created at multiple scales of image resolution, filtered, and then combined to produce a master saliency map. The performance of saliency models is assessed by comparing the saliency maps produced for a range of stimuli to eye tracking data obtained from human viewers looking at the same stimuli.

The MIT Saliency Benchmark project [7] keeps a running scoreboard for author-submitted models, showing how well they predict human fixations on benchmark image sets. The project includes two sets of benchmark images and corresponding fixation data recorded from human viewers. The project has also established eight metrics for assessing the match between saliency and fixation maps [8]. A full discussion of each metric is outside of the scope of this paper (see [8,18] for more detailed descriptions), but each metric is briefly described below.

Three of the eight metrics are location-based, meaning that they assess how well saliency maps predict the location of human fixations in an image. All three of the location metrics are based on the concept from signal detection theory of the Area under the Receiver Operating Characteristic (ROC) Curve, or AUC. The three variants of this approach are AUC-Judd, AUC-Borji, and shuffled AUC (sAUC). Scores range from 0 to 1 with 1 being the optimal score and 0.5 representing chance performance. The key differences between these three metrics lie in how they calculate true and false positives. For example, AUC-Borji uses a uniform random sample, while the sAUC, which was developed specifically for assessing saliency models, samples in a way that penalizes models that are biased toward the center of the image [8].

Four metrics are based on comparisons of the distribution of fixations across an image to the distribution of saliency in a saliency map. These metrics are called the similarity metric (SIM), Earth Mover's Distance (EMD), Pearson's Correlation Coefficient (CC), and Kullback-Leibler divergence (KL). The SIM metric treats the fixation and saliency maps as histograms and assesses their overlap. Scores range from 0 to 1, with 1 indicating perfect overlap. False negatives are highly penalized under the SIM metric. The EMD computes the cost of transforming one map to the other. If two distributions are identical, the EMD is zero, so lower scores represent better performance. CC measures how correlated the two maps are, penalizing false negatives and false positives equally. A score of 1 represents a near-perfect correlation between the saliency and fixation maps. KL is an information theoretic measure that assesses the information lost when the saliency map is used to approximate the fixation map. A score of zero is optimal, so lower scores represent better performance for the saliency map. The KL metric is particularly sensitive to zero values, so sparse saliency maps are penalized with high KL scores [8,18].

Finally, the Normalized Scanpath Saliency (NSS) is a value-based metric. It standardizes the saliency map and then computes the average saliency at locations that were fixated. When the NSS score is greater than 1, that indicates that the fixated locations had significantly higher saliency than other locations in the image [8,18].

The visual saliency modelling community has not settled on any single metric for evaluating model performance. We feel it is important to consider at least one metric from each category (value, location, distribution) because corner cases may be easier to identify when comparing results from metrics in different categories. For consistency with prior publications, and in hope of compatibility with future investigations, we provide results for all of the eight metrics in the evaluations discussed below.

Saliency models are generally trained and tested using images of natural scenes. One of the two sets of benchmark images provided by the MIT Saliency Benchmark, the MIT300 set, consists of 300 images of indoor and outdoor scenes. The other dataset, CAT2000, consists of 2000 training and 2000 test images organized into 20 categories. Of the 20 categories, 15 are comprised of images of natural scenes. These are either photographs or manipulations of photographs, such as inverted or low resolution images. The remaining five categories contain images that are more abstract, such as cartoons, sketches, and fractals.

In a prior study [18], we sought to assess the performance of existing visual saliency models on data visualizations, a category that is not represented in the CAT2000 benchmark. We selected three saliency models that spanned a range of performance on the CAT2000 benchmark: the Itti, Koch and Niebur model [25], the Boolean Map Based Saliency model (BMS) [48], and the Ensembles of Deep Networks Model (eDN) [45]. We measured the performance of each of the selected models on a set of 184 data visualizations drawn from the Massachusetts (Massive) Visualization Data Set (MASSVIS) [6]. These were common types of data visualizations (bar charts, pie charts, etc.) that had corresponding eye movement data from human viewers. For each model, saliency maps were generated for each visualization and compared to the fixation maps using the eight metrics discussed earlier.

This analysis found that all three saliency models generally performed worse on the visualizations than on the images from the CAT2000 data set. The BMS model, which is one of the highest performers on the CAT2000 benchmark, performed significantly worse on data visualizations relative to the CAT2000 images for 6 of the 8 evaluation metrics. The eDN model had significantly worse performance according to five of the eight metrics. Interestingly, the Itti model, which has the lowest average performance of these three models on the CAT2000 set, performed best on the data visualizations. However, it still performed significantly worse on data visualizations than on the CAT2000 images according to four of the eight metrics.

A simple example of the models' underperformance on visualizations is shown in Figure 1, which provides one example from the MASSVIS set with corresponding fixation and saliency maps. Note that most of the fixations (Panel B) were devoted to the text labels for the bar graph. In contrast, the three saliency models tend to predict that viewers will fixate on the bars themselves due to their high

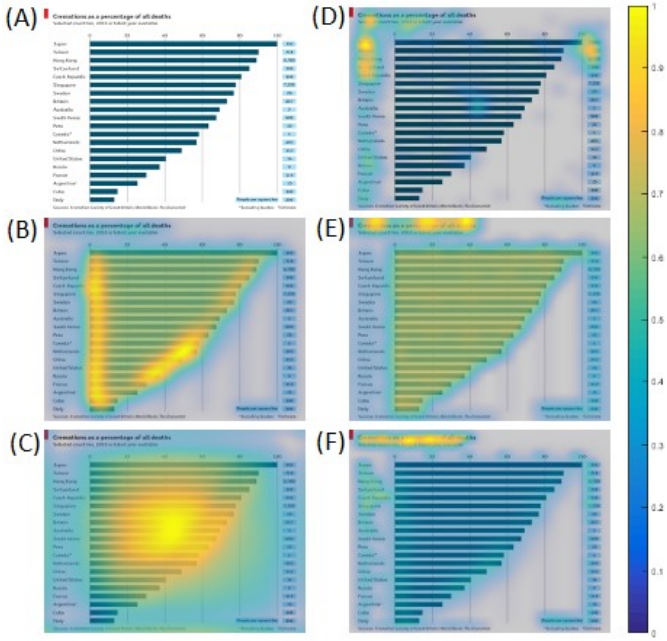



Figure 1: Fixation map and saliency maps generated by different models for an image from the MASSVIS set. (A) the original data visualization; (B) fixation map from Borkin et al. [5]; (C) Itti model; (D) BMS model; (E) eDN model; and (F) DVS model.

contrast, sharp edges, and central location in the image. The reasons for this mismatch are outlined in more detail below.

## 2 Differences in Visual Properties of Data Visualizations and Natural Scenes

It is clear from the analysis outlined above that existing visual saliency models are inadequate for predicting where people will look in abstract data visualizations. Models that generally perform quite well on natural scenes, and even somewhat abstract imagery such as cartoons, performed significantly worse on common types of data visualizations. We hypothesize that the reason for this poor performance is that the spatial scales and visual features used by the saliency models are inadequate for data visualizations.

### 2.1 Spatial Scales

Each of the models discussed above (Itti, eDN and BMS) follows a common approach. First, for each type of visual feature used by the model, "interestingness" maps (or "conspicuity maps," after Itti et al. [25]) are computed at one or more resolutions. Second, the individual feature maps are combined into an overall attention map and then into a saliency map.

As an example, the Itti model operates on multiple spatial scales by constructing a Gaussian pyramid from the input image. At each level of the pyramid, a Gaussian smoothing function is applied and the image is subsampled by a factor of two, creating a smaller, smoothed version of the image, as shown in Figure 2. A feature map is computed for each level, and then the feature maps are compared across levels of the Gaussian pyramid. Image regions with the greatest difference in feature values across scales are assigned higher saliency values than regions with smaller differences across scales. This comparison process is the model's implementation of the center-surround neural activation properties of the human visual system.

Although this approach works relatively well for natural scenes, the spatial properties of data visualizations are quite different. Many of the elements in data visualizations (glyphs, lines, text) are quite small, and visualizations are likely to have a higher proportion of small but important variations than natural scenes. The smoothing and subsampling process results in the loss of these small details. For example, text becomes blurry at the first level of smoothing, leading

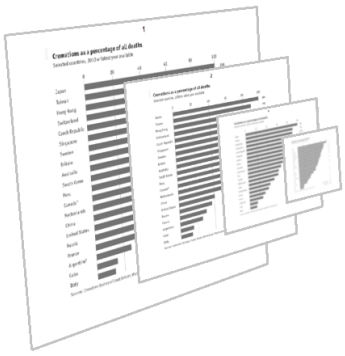


Figure 2: Example of a Gaussian pyramid with four levels of smoothing and resizing.

to minimal differences between the levels of the Gaussian pyramid when the visual features of the text are compared across scales. This results in low saliency values for text even though text typically receives a high proportion of fixations [37].

Another problematic aspect of the existing saliency models is that many of them resize the input image to a standard size as their first step. For example, the BMS model begins by resizing the input to be exactly 600 pixels wide. Similarly, the reference implementation of the eDN model resizes its input to a resolution of 512x384. While this makes the computation go quickly, it also tends to blur text into unrecognizability and obliterate fine contours completely. This is a particular problem for visualizations since the meaningful elements of many data representations (line charts, box charts, some geographic maps and weather diagrams) are nothing but fine contours.

### 2.2 Visual Features

While the way in which the models combine their feature maps is fundamentally similar, they differ in terms of the specific visual features used to create the feature maps. The Itti model computes center-surround operations on intensity, orientation and color channels and combines them to create the attention map. It computes four color maps (red, green, blue and yellow) using RGB pixel values. The eDN model uses a support vector machine trained over many randomly constructed hierarchical features [42]. These features operate variously on RGB, YUV and grayscale images. The BMS model uses exactly one feature – connected regions. It computes these regions at multiple intensity thresholds using the channels of the CIE LAB color space.

#### 2.2.1 Color

Since all three models compute some or all of their features over color channels, we believe that the color space chosen for these computations is particularly important. In our assessment of the three saliency models using the MASSVIS images, we noted that the models often assigned low saliency values to bright red regions, causing discrepancies between the saliency maps and the map of human fixations. We believe that this mismatch is driven by the fact that human color perception is very different from the way colors are created on paper or on an electronic display. This difference manifests in two ways. First, color spaces such as RGB or CMYK that are defined by the properties of an output device are perceptually non-uniform. That is, adding 0.1 to the red component of a color produces a larger perceived difference for some colors than for others. Second, the different "channels" of human color perception are not independent as they are in the case of display primaries. That is, adding redness while keeping luminance constant may change perceived luminance.

The YUV color space uses a luminance + chrominance representation of color that it is designed to permit efficient compression while minimizing artifacts. From the perspective of perceptual uniformity, YUV is an improvement over RGB but still leaves much to be desired. In order to do color arithmetic in a way

that yields perceptually comparable results, it is advisable to work in a color space like CIE XYZ or CIE LAB [14]. The XYZ model operates with the tristimulus values obtained from the color-sensitive cones in the retina. The LAB model transforms these into a luminance channel (L) and two color-opponent channels (A and B) that agree with current thinking about the way color is processed in the brain. The LAB model has the additional advantage of being perceptually uniform. Adding 0.1 to a color component produces a change that appears to the observer to be of the same magnitude regardless of where it is in the color space. As a result, feature maps computed over different channels in the color space have values that can be meaningfully compared with one another.

#### 2.2.2 White Space

A crucial difference between visualizations and natural scenes is the presence of white space. The real world is cluttered and natural scenes tend to have information (in the Shannon sense) absolutely everywhere. Synthetic scenes do not: they often contain large areas of uniform, untextured color. Some of these may be objects, but some are simply blank areas. Distinguishing between the two is a challenge. In either case, feature-based saliency models may have trouble "seeing" these regions since they will only be detectible a very coarse scale.

The spatial distribution of figures relative to the background is also quite different for abstract data representations than for physical objects. Many saliency models use a center weighting. This works well for photographs, where objects of interest are often centered. However, it may not be appropriate for visualizations, where meaningful information can appear in any spatial location and is often deliberately distributed across the entire image.

#### 2.2.3 Text

As mentioned above, text in data visualizations receives a great deal of attention from viewers. In prior work, we have found that people viewing data visualizations while performing memory or free viewing tasks devote a disproportionate amount of attention to regions containing text. For example, in one dataset, an average of 60% of the participants' fixations fell in regions containing text, relative to 30% in regions containing visual representations of data [37]. In general, participants were highly likely to view regions containing text and to view them relatively early in the trial.

There are several causes for the high proportion of fixations devoted to text in visualizations. In general, literate people's attention is automatically drawn to text [28,33,35]. In data visualizations, text often provides context and details that are necessary for understanding the data. For example, our prior work found that participants are likely to refer to text-containing regions such as the legend and data labels multiple times as they view the visualization [37]. Finally, reading text requires numerous fixations. Under normal conditions, the estimated visual span for reading is about 10 letters [31]; words presented in peripheral vision cannot be resolved due to low visual acuity and crowding.

While text draws attention and necessitates many fixations, it is not included as a feature in most saliency models. The models are tailored to and/or trained on images of natural scenes, which rarely contain text. Our analysis of the performance of existing saliency models on data visualizations indicates that assigning appropriate levels of saliency to text is one of the key areas in which their performance could be improved.

## 3 The Data Visualization Saliency Model

Existing saliency models fall short for data visualizations, but our analysis of several models revealed concrete steps that can be taken to adapt them to this domain. We have developed the Data Visualization Saliency (DVS) model[1], which builds on the strengths of existing models while extending their capabilities to account for the visual features and spatial scales that are common in data visualizations. The two primary components of the current implementation of the model are a modified version of the Itti model and a text recognizer, which allows us to detect one of the key features of visualizations that is missed by current models. The DVS model combines the outputs of the modified Itti model and a text map to produce saliency maps that are specialized for data visualizations.

### 3.1 Modified Itti Model

We took as a starting point the Itti, Koch and Niebur saliency model [25] as implemented in the Graph Based Visual Saliency (GBVS) toolbox [20,21]. Of the existing models that were tested with data visualizations, this model had the highest performance [18]. The authors of the GBVS saliency model note that the original Itti model uses a simple color opponency representation based on RGB values. As discussed above, using the RGB color space is suboptimal, particularly in the case of data visualizations, where colors are chosen deliberately by a designer. To better approximate human visual perception, we modified the original algorithm by transforming the representation of the input images into CIE LAB color space. This change is likely to improve the model's performance for all types of imagery, but it is particularly important for visualizations, in which colors are deliberately selected to convey information.

### 3.2 Text Saliency Map

As discussed above, viewers devote a great deal of attention to text in data visualizations, yet text is not highlighted in existing saliency models. Although text regions often have high contrast, they tend to be small. The high-frequency details of text are lost when an input image is resized or smoothed. This leads to few differences across the levels of the Gaussian pyramid, and the text regions are not identified as being salient. To account for viewers' tendency to fixate on text in visualizations, we developed a text saliency model that could be combined with the modified Itti model. Attention to text is primarily driven by top-down visual attention, since people expect text to contain meaningful information. By incorporating this feature into our model, we are taking a step towards a saliency model that takes both bottom-up and top-down attention into account.

Our goal was to build an algorithm that computes the likelihood of belonging to a text region for each pixel of an input visualization image. Text detection is a popular challenge in the computer vision literature, and numerous successful models and algorithms have been developed in this domain. Detecting text in visualizations is a relatively easy task compared to detecting text from photos of real-world scenes. The method we detail below is essentially a combination of various classic text detection techniques. However, instead of producing a binary output, like traditional text detection algorithms, this method produces a continuous, probabilistic output that can be incorporated into a saliency map.

We used a common approach in the text detection literature, which is to extract Maximally Stable Extremal Regions (MSER) [36] as candidate text regions, and then to apply various text-diagnostic features to filter out the non-text candidates (e.g., [11,17,40]). The MSER algorithm detects connected, homogeneous ("maximally stable") regions of pixels. Because text almost always has uniform color and each letter in English is connected (in the sense that each "stroke" is connected to all other strokes in the same letter), English letters should be detected as MSER regions (i.e., the miss rate should be very low).

In order to exclude MSER regions that are not text, all detected MSER regions went through a filtering process based on simple properties of these regions, such as aspect ratio [11], Euler number [17,40], and solidity [17]. As an example, for most fonts of English letters and Arabic numerals, the height-to-width ratio should be less than 4 and greater than 1/3, so the aspect ratio of the bounding box of

[1] Available at: https://github.com/mjhaass/DataVisSaliency.git

MSER regions was restricted to this range [11]. Finally, the data was filtered based on stroke width variation [17,32]. The variability of each MSER component's stroke width was compared to its mean stroke width. If the relative variability was too large, the region was filtered out (since letters and digits have relatively small stroke width variations).

After the above filtering, the remaining MSER regions had a relatively high likelihood of being letters or digits. In order to quantify this likelihood, we computed three text-diagnostic edge features on these regions (using simplified versions of the algorithms proposed by [34]). We took the bounding box of each MSER region and computed these features on the image patch defined by the bounding box. The three feature values were then summed together to form the raw "text saliency" score.

The first feature was based on the magnitude of the image gradient. For each image patch (i.e., each MSER region), the image gradient was computed on the grayscale transformation of the original colored patch. The mean gradient magnitude $\mu(G)$ and the standard deviation $\sigma(G)$ of gradient magnitude were computed with $P$ as a scaling constant:

$$F_1 = P\frac{\mu(G)}{\sigma(G)} \tag{1}$$

This feature is akin to a signal-to-noise ratio. In most scenarios, text strokes appear on a highly uniform background; the variability of the gradient magnitude is low but the text edges lead to high gradient magnitudes. This ratio should be high when the image patch contains text.

The remaining features were based on the edges in an image patch. For each MSER region, the Canny edge detection algorithm [9] was used to compute an "edge image" for each color channel of the image patch as represented in the CIE LAB color space.

The second feature attempts to capture a specific topological characteristic of text. Most text characters have either multiple strokes that intersect each other or curved strokes so that a vertical or horizontal "scan line" may cross the character body more than once. Since each stroke produces two edges, such "scan lines" will very likely cross the edges of the character more than twice. Therefore, the frequency of multiple-crossing by a scan line that scans horizontally and vertically is diagnostic of text. The higher the frequency, the more likely the image patch contains text. Formally, this feature can be given as

$$F_2 = Q^{\left(\sum_{i=1}^{H} f(cn_i)+\sum_{j=1}^{W} f(cn_j)\right)/(W+H)} \tag{2}$$

where $W$ and $H$ are the width and height of the image patch in pixels, $cn_i$ and $cn_j$ denote the number of crossings for a specific scan line (vertical and horizontal respectively) and the edges in the image patch, and $f(x)$ is a function that returns 1 when $x$ is larger than 2 and 0 when $x$ is equal to or less than 2. The constant $Q$ is for scaling and weighting purposes. Using an exponential function with base $Q$ increases the feature's sensitivity to higher multiple-crossing event counts and reduces sensitivity to small counts (which can occur randomly in non-text regions).

The third feature was based on a more straightforward characteristic. Text strokes usually produce two parallel edges, so that the number of crossings between a vertical or horizontal scan line and the text edges is often an even number. Hence the third feature can be defined similarly to the second one:

$$F_3 = R^{\left(\sum_{i=1}^{H} g(cn_i)+\sum_{j=1}^{W} g(cn_j)\right)/(W+H)} \tag{3}$$

where $g(x)$ returns 1 if $x$ is an even number and 0 if it's odd. In the current implementation, the values of the scaling constants are $P = 2.5$, $Q = 4$, $R = 1.22$.

The text-specific feature values were normalized, combined, and treated as an index of probability of text in each region. The combined value of the three features was assigned to the pixel at the center of the region. This procedure was computed at different scales on the original image in order to enhance the method's sensitivity to smaller and larger fonts. The text saliency indices computed at each scale were re-scaled to the original image size and then combined by averaging. This raw text saliency map was then processed with Gaussian smoothing to simulate the randomness in the exact locations of human fixations.

### 3.3 Linear Combinations of the Model Components

Because there is insufficient data to inform how to best combine the text saliency map and the modified Itti saliency map, we opted for the simplest approach: a linear combination. Formally, the DVS model's saliency map S for a given visualization is computed as follows:

$$S = \frac{(I+w*T)}{(1+w)} \tag{4}$$

where $I$ is the saliency map given by the modified Itti saliency model, and $T$ is the text saliency map. The parameter $w$ determines the relative weight between $I$ and $T$. Both $I$ and $T$ are linearly scaled to have values ranging from 0 to 1 before combination. The denominator, $(1 + w)$, produces a weighted average to maintain the overall saliency scaling from 0 to 1. Thus, for each data visualization image, a series of saliency maps based on a series of weight values can be generated. In order to choose an appropriate weight for the text saliency map, we systematically manipulated linear combinations of $I$ and $T$ and compared the resulting saliency maps to eye tracking data from the MASSVIS project [5]. The MASSVIS data set provides 393 data visualization images and corresponding fixation data. Thirty-three participants viewed the images while trying to memorize them for a later test. One visualization was excluded from our evaluation because it had an irregular size (less than 128 pixels wide) that is incompatible with the Itti saliency model. Thus, saliency maps and performance metrics were computed on the remaining 392 images.

We were primarily interested in how the average value for each of the eight MIT Saliency Benchmark evaluation metrics changed as a function of relative weight w between the modified Itti saliency map $I$ and the text saliency map $T$. When $w = 0$, the saliency map $S$ is just the modified Itti map; similarly, when $w \to \infty$, $S$ is equivalent to the text saliency map $T$. If the bottom-up saliency component captured by the modified Itti map $I$ and the text-directed attention captured by $T$ do complement one another, at some nonzero value of $w$, the combined map $S$ should provide higher performance than either $I$ or $T$. In other words, the performance-relative weight function should have a maximum point. Because of the differences in the nature of these metrics, we expect these functions to have different maximum points. Our goal was to find a reasonably good estimate of the window of $w$ values in which the function reaches maximum for each of the eight metrics. Figure S1 in the Supplemental Materials plots each metric as a function of the weight parameter.

Notably, the baseline performance for the text saliency model was better than the baseline performance of the modified Itti model for six of the eight metrics (the SIM and KL metrics were the exceptions, likely because the text saliency maps include large regions that contain only zeros, and both of these metrics heavily penalize false negatives). The preference for the text saliency model is consistent with prior analyses showing that viewers disproportionately devote their attention to the text in the MASSVIS images [37]. Modelling only the text regions is a reasonable approximation for where people look in this particular data set and task. However, across all eight metrics, the linear combination of the modified Itti model and the text saliency model produced significantly higher matches to the human fixation data than either model alone.

The weight functions for each metric exhibit different shapes, reaching their maxima at different weight values. This aspect of the data was expected and supports the assertion that the eight metrics

emphasize different aspects of the performance of a saliency model. There is no objectively optimal choice of the text saliency map weight, since no unique weight value optimizes all metrics of performance. In our experience, the choice of weighting factor typically causes performance results to fall into one of three categories; *under fit*, where performance increases proportionally to the weighting factor, *acceptable*, where the performance is stable, or changes very slowly with changing weighting factor, and *over fit*, where performance may increase, but the gain on a given test case is likely not to transfer to another test case. Figure S1 shows that at least four of the performance metrics are approaching an asymptotic limit as the weight factor value approaches 2. To reduce the risk of over fitting, we chose to use a weight of 2 in the following analyses. Users of the DVS model can easily adjust this weight, if desired.

Figure 3 shows a representative example of the differences between the DVS model and the original Itti model. Additional examples are provided in the Supplemental Materials. The top panel of Figure 3 shows a data visualization from the MASSVIS set with overlaid fixation data (A). The remaining panels show the saliency maps produced by the original Itti map (B), the modified Itti map (C), the text saliency map (D), and the final, weighted DVS saliency map (E). Finally, the bottom panel (F) shows the DVS map overlaid on the original image, using the same color scale as the fixation map, allowing for a visual comparison of the two. Note that the original Itti map identifies the lower portion of the bar chart as the most salient region. The differences between the original Itti map and the map with the modified color space are subtle, but the modified model appears to do a better job of picking out the line graphs. The text saliency map correctly identifies all of the text regions in the image, but also has a few false alarms to features in the data, such as the data points on the line graphs. The DVS saliency map indicates that the title is highly salient, as is the lower part of the chart and the labels at the bottom of the chart. This corresponds well to the actual distribution of viewers' fixations.

### 3.4 Comparing the DVS Model to Existing Saliency Models

Once the weights in the DVS model had been optimized, the performance of the final model was compared to the original Itti model (as implemented in the GBVS toolbox), the BMS model, the eDN model, and to the text saliency maps alone. All of the models were used to generate saliency maps for 392 data visualizations from the MASSVIS dataset that had corresponding eye tracking data (as before, one visualization was excluded because its dimensions were incompatible with the Itti model). The saliency maps were compared to the eye tracking data using the eight metrics that are used by the MIT Saliency Benchmark. A one-way ANOVA was run for each metric, showing that there was a significant difference in the performance of the five models on all eight metrics (all $F$s > 44.69, all $p$s < 0.001).

Table 1 shows the percentage of improvement for the final, weighted DVS model relative to the Itti, BMS, eDN, and text saliency models on all eight metrics. The DVS model offered a substantial improvement in performance over the other models. Since the DVS model is based on the Itti model, we paid particular attention to how the components of the DVS model performed relative to the original Itti model. Figure 4 shows the effect size, using Glass's delta, for the improvement in performance for the text saliency maps and the final DVS model relative to the original Itti model. Notably, for all of the metrics other than EMD, the improvement in performance over the original Itti model was larger than one standard deviation. Performance also improved for the EMD metric, but the magnitude of the improvement was smaller. Finally, we used paired t-tests to assess whether or not the DVS model, as implemented with a weighting of 2, performed better than the text saliency maps alone. The KL metric was excluded from this analysis because its high sensitivity to zero values produced abnormally large scores for the text saliency maps. The DVS model performed significantly better than the text only

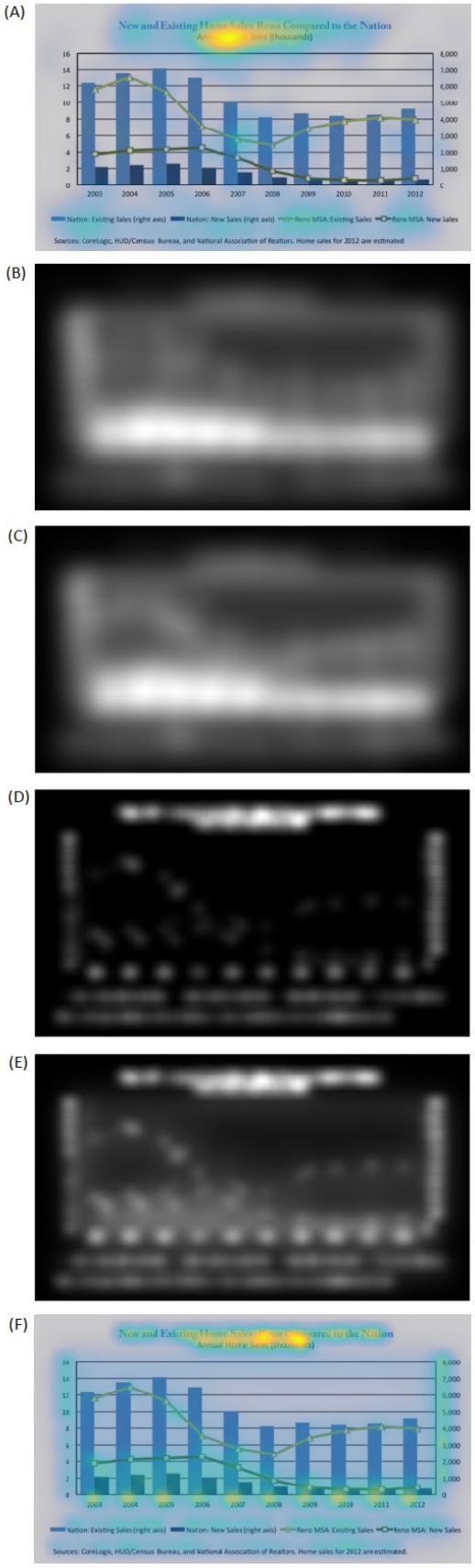
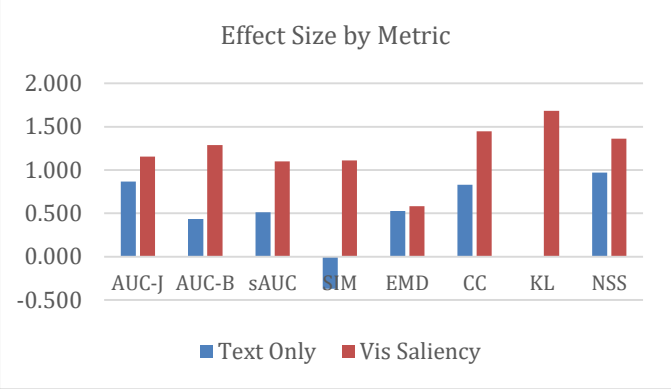



Figure 3. (A) An image from the MASSVIS set overlaid with fixation data and saliency maps produced by the original Itti (B), modified Itti (C), text saliency (D), and DVS (E) saliency models, with the DVS map overlaid on the original image in (F).

model as measured by six of the seven metrics (all $t$s > 2.04, all $p$s < 0.02). The only exception was the EMD metric ($t(391) = 0.65$, $p = 0.26$). In this case, the scores for the text only and DVS models were nearly identical.

Table 1. Percentage Improvement for the DVS Model Relative to the Itti, BMS, eDN, and Text-Only Models.

| | | Itti | BMS | eDN | Text |
|---|---|---|---|---|---|
| **Location Metrics** | **AUC-J** | 9% | 11% | 24% | 2% |
| | **AUC-B** | 9% | 12% | 22% | 5% |
| | **sAUC** | 9% | 11% | 21% | 4% |
| **Distribution Metrics** | **SIM** | 9% | 14% | 18% | 15% |
| | **EMD** | 18% | 21% | 26% | -1% |
| | **CC** | 41% | 70% | 133% | 5% |
| | **KL** | 20% | 37% | 33% | -- |
| **Value Metric** | **NSS** | 55% | 82% | 176% | 2% |




.

Figure 4. Effect size, using Glass's delta, of the improvement due to using the DVS model for all eight metrics.

## 4 Testing the DVS Model's Performance

While the DVS model outperformed the Itti model in our initial assessment, there are several factors that limit our ability to generalize these findings. First, the MASSVIS data were collected in the context of a memory study, which might bias participants to focus more on the text in the visualizations. In addition, participants in the MASSVIS study viewed the images for 10 seconds, which is a longer duration than is typically used for comparing fixation data to saliency maps. For example, the widely-used eye tracking data sets provided by the MIT Saliency Benchmark had images that were presented for three seconds (MIT300) [27] or for five seconds (CAT2000) [4].

To get a broader understanding of the performance of the DVS model relative to existing saliency models, we used an additional data set to compare the performance of the DVS, Itti, BMS, and eDN models. This data set [37] consisted of eye tracking data collected from 30 participants who viewed four types of stimuli. As in the CAT2000 dataset, the participants viewed each stimulus for five seconds under free viewing conditions. The stimuli were presented in four counterbalanced blocks. One block contained 35 data visualizations from the MASSVIS dataset. Another contained 27 newly-generated, simple data visualizations that contained relatively little text. This set contained three visualizations of each of the following types: bar charts, box plots, bubble plots, column charts, line plots, parallel coordinates plots, pie charts, scatter plots, and violin charts. The other two blocks contained stimuli from the CAT2000 dataset [4] that were selected for their visualization-like properties. One block contained 30 line drawings and the other contained 16 images of fractals. These materials were chosen because they have already been incorporated into assessments of visual saliency models, yet like data visualizations, they have visual properties that differ from those of natural scenes. The line drawings have the same overall spatial layouts as natural scenes, but no colors and many fine contours that may be lost when the images are smoothed and resized by the saliency models. The fractals have very different spatial properties and color palettes than natural scenes, with vivid colors and shapes that fill the entire frame. Like data visualizations, they are abstract and computer-generated.

For each subset of stimuli, we assessed the match between the human fixation data collected by Matzen and colleagues [37] to the saliency maps produced by the DVS, Itti, BMS and eDN models using the eight MIT Benchmark metrics. In addition, as a point of reference, we compared the fixation data across experiments. For the MASSVIS stimuli, fixations were compared across the Matzen and colleagues [37] dataset and the original MASSVIS study [5]. For the fractal and line drawing stimuli, the fixation data was compared to the MIT Saliency Benchmark fixation data [4,7]. Although different groups of participants viewed the stimuli in the various experiments, and in the case of the MASSVIS data, the participants were performing a different task, we would expect to see the highest scores on the eight metrics when comparing one set of human fixations to another. If the models can accurately predict where viewers will look in data visualizations, their performance should approach the level of agreement between the two sets of fixation maps.

The results of the analysis for the line drawing stimuli are shown in Table S1 in the Supplemental Materials. These stimuli are most similar to natural scenes in terms of their spatial properties. As expected, the comparison between the two sets of fixation data had the best similarity scores for most of the metrics (six of the eight). When comparing the performance of the four models against the Matzen and colleagues [37] fixation data, the eDN model had the best scores for four of the eight metrics, the Itti model had the best scores on three of the metrics, and the DVS model had the best score on one metric, the sAUC.

The results of the analysis for the fractal stimuli are shown in Table S2 in the Supplemental Materials. These stimuli are somewhat of an intermediate point between natural scenes and data visualizations. They are computer generated and do not have naturalistic colors or spatial layouts, yet they do not contain text and their visual elements are not intended to convey specific information to the viewer. For these stimuli, the comparison of the two sets of fixation data had the best similarity scores for all eight metrics. When comparing the models to the fixation data, the eDN model had the best scores for six metrics and the DVS model had the best scores for two of the metrics.

The results of the analysis for the simple data visualizations are shown in Table S3 in the Supplemental Materials. When the four sets of saliency maps were compared to the fixation data, the DVS model had the best scores for seven of the eight metrics. The Itti model had the best score on the AUC-Borji metric.

The results of the analysis for the MASSVIS stimuli are shown in Table S4 in the Supplemental Materials. Once again, the comparison of the two sets of fixation data led to the best similarity scores for all eight metrics. When comparing the models to the fixation data, the DVS model had the best scores for all eight metrics.

To test whether or not the DVS model performed significantly better than the Itti, BMS and eDN models for data visualizations, the two sets of visualizations were combined. One-way ANOVAs were conducted for each of the eight metrics. These ANOVAs showed that there was a significant difference in performance across models for all eight metrics (all $F$s > 22.37, all $p$s < 0.001). Post-hoc t-tests showed that the DVS model's scores were better than the other models' scores for seven of the eight metrics (all $t$s > 3.74, all $p$s < 0.001). The exception was the AUC-Borji metric. According to this metric, the DVS model performed significantly better than the BMS ($t(61) = 6.50$, $p < 0.001$) and eDN ($t(61) = 9.34$, $p < 0.001$) models, but not the Itti model ($t(61) = 1.20$, $p = 0.12$).

### 4.1 Discussion

Our comparison of the Data Visualization Saliency model to the Itti, BMS, and eDN models found that the eDN model was generally the

highest performer for line drawings, images that are somewhat abstract, but that share the spatial properties of natural scenes. This is consistent with the eDN model's overall high performance on the MIT Saliency Benchmark, the source from which the line drawing stimuli were taken. Similarly, the eDN model was also the best performer for fractal stimuli, which were also drawn from the MIT Saliency Benchmark set. We observed that the eDN model tends to produce saliency maps with a pronounced center weighting. This aligns well to the fixation maps for the fractal stimuli, where participants tended to fixate most on the center of the images.

For the line drawing and fractal stimuli, the DVS model's performance was typically similar to, or slightly better than, that of the Itti model, the model on which it is based. This indicates that our changes to the Itti model's color maps and the addition of the text saliency maps does not hinder the model's performance on stimuli that are not data visualizations. We anticipate that this would be true for images of natural scenes as well. The improved color map provides small improvements to performance, while the text saliency map contains only zero values in a scene that has no text, so it does not impact the final DVS map for such scenes.

Since our focus is on developing a saliency model that can be used as an evaluation tool for data visualizations, those stimuli provide the most important test of the model's performance. Our test set included two types of data visualization stimuli: simple visualizations that contained minimal text, no contextual information, and no "chart junk," and in-the-wild visualizations culled from publications, which typically contained explanatory text, source information, and graphical elements chosen for aesthetic or branding reasons. For the simpler data visualizations, the DVS model had the best performance according to seven of the eight metrics, and for the more complex visualizations, it had the best scores for all eight metrics. These results show that modifying the color map of the Itti model and adding a new visual feature (text saliency) led to significantly better performance on data visualizations.

For the MASSVIS stimuli, we were able to compare fixation data recorded from two different populations of participants in two different experimental contexts [5,37]. This comparison is in some sense a benchmark for model performance. If the models can accurately predict human fixations, their performance should approach the level of similarity obtained by comparing two sets of fixation data. The DVS model's scores were the closest to the scores for the fixation-to-fixation comparison for all eight metrics, and for the sAUC and KL metrics, paired t-tests showed that there was not a significant difference between the two scores ($t(34) = 0.01$ for sAUC, $t(34) = 0.04$ for KL).

## 5 Applying the DVS Model

Our results indicate that, of the models tested, the saliency maps produced by the DVS model were the best match to maps of human fixations, approaching the level of fixation-to-fixation comparisons in some cases. This suggests that the DVS saliency maps provide a reasonable approximation of which regions of a visualization are most likely to draw the viewer's attention.

As described above, this provides a useful evaluation metric for visualization designers. Ideally, the most important information in a visualization will also be highly salient [26,38]. Jänicke and Chen [26] illustrated this approach by using the Itti model as an evaluation tool. They compared saliency maps generated by the Itti model to a "relevancy map" defined by the visualization designer. They suggest that this comparison can be used to evaluate different visualization techniques or candidate visualizations in order to choose the one that most effectively highlights the important information.

The DVS model represents an improvement over the Itti model, but it can be used in a similar manner to evaluate visualizations. For example, the DVS saliency map in Figure 3 shows that the viewer's attention is most likely to be drawn to the text, the dark blue bars, and the tops of the light blue bars upon his or her initial viewing of the visualization. However, suppose that the visualization designer knows that the data represented by the line graphs is particularly important. The DVS saliency map provides a quick and easy way to assess whether or not this visualization will draw attention that data. In this example, the line graphs are not very salient, so the match between the importance of the data (i.e., top-down goals) and its salience (i.e., bottom-up attention) is poor. Armed with this information, the designer can try other variants of the visualization or other visualization techniques in order to select one that makes the most important information more salient.

The simplest way to evaluate a visualization using a saliency model is to take a qualitative approach. A designer can generate saliency maps for a set of visualizations and compare them visually, identifying the options that have a good distribution of saliency (as defined by the designer's goals). However, the saliency maps can also be used in a quantitative fashion. As suggested by Jänicke and Chen [26], designers could define a relevancy map and assess the match between the relevancy and the saliency maps. This assessment could be done categorically, as in their paper, or it could be done using one or more of the eight metrics that are commonly used to assess saliency maps. If only one is used, we propose that the value-based NSS metric would be the most appropriate for this type of comparison. If the designer assigns a relevancy value to each region of a visualization, the NSS metric can be used to assess the match between the relevancy values and the saliency values at each location. One prior study [23] has used the NSS metric to compare fixation data to important features in 2D flow visualizations, so there is some precedent for using this particular metric in the context of evaluating visualization techniques.

Another approach to quantitative assessment is to define regions of interest that outline the most important features in the data. After generating a saliency map, a designer could assess what percentage of the saliency falls within the regions of interest. This provides a simple numerical assessment of the match between the importance of the data and its saliency. To aid in evaluation, we have implemented this feature in the DVS model. A user can input the coordinates of a polygon describing a region of interest, and the model will provide the percentage of visual saliency, normalized for overall area, that falls within that region.

## 6 General Discussion

Visual saliency models have been the focus of a great deal of research in the cognitive science and computer vision communities because mimicking human visual attention has numerous applications, including image compression, image segmentation, object recognition, visual tracking, and image quality assessment [38,45,49]. Visual saliency maps could also play a role in evaluating data visualizations by allowing designers to determine whether or not a particular visualization draws the viewer's attention as intended. Since saliency models are inspired by the properties of the human visual system, the same system that is used to convey information in data visualizations, these models have the potential to serve as a simple and general evaluation tool.

While visual saliency models have a great deal of potential as an evaluation metric, prior evaluations have shown that existing saliency models consistently underperform on data visualizations, often failing altogether [18]. The models that perform best with natural scenes perform worst on data visualizations, and vice versa. Through assessments of three saliency models that generally perform well for natural scenes, we found that the spatial scales and visual features used by the existing saliency models are inadequate for data visualizations. Two particularly problematic areas were color models and text. The existing models perform operations using color spaces that do not correspond well to human perception of color. And while text draws a great deal of human attention, it is typically missed by saliency models due to its small spatial extent and high-frequency variation. Color and text are both very important features of data visualizations, chosen by designers to convey specific information to viewers. Thus,

we chose to focus on these two areas in order to develop a saliency model that makes more accurate predictions of where viewers look in data visualizations.

We based the Data Visualization Saliency (DVS) model on the Itti model, which performed better than other existing saliency models on data visualizations. We modified the Itti model to use the CIE LAB color space, which is more representative of human color perception, and added a model of text saliency. We used a linear combination to incorporate the text saliency maps into the modified Itti model, and optimized the weighting of each component by testing the model against the stimuli in the MASSVIS dataset. To assess the performance of the final, weighted model, we compared its performance to the original Itti, BMS and eDN models using a set of fixation data obtained from participants viewing line drawings, fractals, and data visualizations [37]. We found that the DVS model's performance was comparable to the original Itti model's performance on the line drawing and fractal stimuli, and that it performed significantly better than the other models for data visualizations.

We suggest that the resulting model could be a simple and useful evaluation tool, which visualization designers can use to compare candidate designs in either a qualitative or quantitative manner. This approach is broadly applicable, but it may be particularly relevant to the evaluation of emphasis effects. There are numerous techniques that have been developed to emphasize subsets of the data in a visualization (see [19] for a review and evaluation framework). Hall and colleagues [19] frame emphasis effects in terms of visual prominence, which is another way of describing visual salience. They discuss intrinsic prominence, driven by the initial process of creating a visual mapping for data, and extrinsic emphasis effects, such as zooming and highlighting, that are used to enhance the prominence of selected features. Saliency maps could be used to evaluate both types of effects and to determine when one type of emphasis overrides the other. An evaluation based on visual saliency is particularly suited to assessing emphasis effects, since many of the features that are commonly used for emphasis (e.g., changes in color or size) are the same features that are used by saliency models.

Evaluations using visual saliency maps are complementary to other evaluation techniques, such as eye tracking. Eye tracking is a useful evaluation tool in its own right, and has been growing in popularity [13,15,16,29,44]. In our prior work with scene-like visualizations, we showed that eye tracking and saliency maps could be used in combination to assess the importance of features in the data and to understand the impact of users' expertise on their attention to those features. This provides information about how the visualization could be modified to better support the users' needs [38]. However, while eye tracking can be very informative, these studies can also be very time consuming and complex. Saliency maps provide a prediction of where users are likely to look without the need for eye tracking, and for many evaluation contexts, this may be sufficient.

### 6.1 Limitations and Future Directions

Although this model has the potential to be a simple and generalizable evaluation metric, there are several limitations to this approach. One important limitation is that the DVS model currently applies only to static images. This is a limitation both because interactions are a key component of many visualizations and because motion is a visual feature that typically captures human attention. In its current implementation, the DVS model can be applied to still images representing different phases of an interactive process, but it cannot capture the interactive component itself. In future work, motion detection algorithms could be incorporated into the model, enabling it to predict which parts of a dynamic scene will draw the viewer's attention most strongly. This would improve the model both in terms of its representation of human visual processing and in terms of its utility as an evaluation tool.

Another limitation is that the current implementation of the model does not change the spatial scales used by the Itti model, although these can also be problematic when applied to visualizations. The model resizes and smooths images, resulting in the loss of fine-grained details that are often very important in data visualizations. In future work, we plan to address these issues by allowing larger input images (limiting the need for resizing) and exploring the effects of changing the scales at which multiresolution differences are calculated.

A limitation of saliency models in general is that they focus on bottom-up visual attention. Bottom-up attention is only part of the picture, and top-down visual attention, driven by the viewer's task, goals, and prior experience, is also of tremendous importance in determining where a person will look in an image or a visualization [22,38,47]. Viewers with different goals may look at completely different parts of the same visualization. The DVS model incorporates one aspect of top-down attention by incorporating attention to text. Small regions of text may not be very salient from a bottom-up perspective, but people look at these regions because they expect them to convey meaningful information. In the future, additional feature detectors could be incorporated into the model to capture common graphical codes that convey semantic information in data visualizations [46], as these would also have high importance from the perspective of top-down attention. The eight evaluation metrics could be used to assess how the performance of the model changes with the addition of each feature.

On the other hand, the addition of more top-down features could quickly reduce the generalizability of the model. Text is unique in some sense because all literate people have extensive experience with processing text, to the point where it becomes automatic and involuntary [28,33,35]. That is not necessarily the case for other features that are used in visualizations. This could lead to differences between users with different levels of experience with the visualization technique or with the domain.

An alternate approach may be to incorporate Gestalt-based features into the model, since many visualization techniques are rooted in Gestalt psychology [46]. Like text comprehension, Gestalt principles reflect general cognitive processes that are not dependent on knowledge of any particular domain. The BMS saliency model relies on the Gestalt principle of figure-ground segregation to identify figures within an image [18,48], so incorporating Gestalt principles into a saliency model is certainly feasible. The BMS model does not perform well for visualizations [18], indicating that this principle alone is not sufficient for our purposes. However, it may be possible to use a similar approach to implement Gestalt-based features within the DVS model. The combination of the modified Itti maps, text saliency maps, and Gestalt-based maps could further improve the model's performance. This is an area that we would like to explore in future research.

Visualizations serve a variety of functions and support a vast range of tasks, so there is an enormous range of factors that might influence the viewer's top-down, goal-oriented processing. The wide range of roles for visualizations is part of what makes evaluation difficult in the first place! Saliency models cannot solve this problem, even with the addition of more features that are inspired by top-down attention. However, despite their imperfections, they can still be a useful tool in a designer's evaluation tool kit. If a designer has a sense of what information is most important from a top-down perspective, she can then assess the visual saliency of her design to determine whether or not the most important features are also salient from a bottom-up perspective. This provides a simple and rapid assessment that can be used in a quantitative or qualitative fashion to inform the visualization's design.


### ACKNOWLEDGEMENTS

This work was supported by the Laboratory Directed Research and Development (LDRD) Program at Sandia National Laboratories. Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

## REFERENCES


[1] J Aloimonos, "Purposive and qualitative active vision," *Proc. 10th International Conference on Pattern Recognition,* pp. 346-360, 1990.

[2] J Atkinson, "The Developing Visual Brain" Oxford University Press, Oxford, UK., 2002.

[3] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 35, pp.185-207, 2013.

[4] A. Borji and L. Itti, "Cat2000: A large scale fixation dataset for boosting saliency research," *CVPR 2015 Workshop on the Future of Datasets*. arXiv preprint arXiv:1505.03581. 2015.

[5] M. A. Borkin, Z. Bylinskii, N. W. Kim, C. M Bainbridge, C.S Yeh, D. Borkin, H. Pfister, and A. Oliva, "Beyond Memorability: Visualization Recognition and Recall," *IEEE Trans. Visualization and Computer Graphics*, vol. 22, pp.519-528, 2016.

[6] M. Borkin, Z. Bylinskii, G. Krzysztof, N. Kim, A. Oliva, and H. Pfister, "Massachusetts (Massive) Visualization Dataset," http://massvis.mit.edu. 2017.

[7] Z. Bylinskii, T. Judd, A. Borji, L. Itti, F. Durand, A. Oliva, A. Torralba, "MIT Saliency Benchmark," http://saliency.mit.edu. 2017.

[8] Z. Bylinskii, T. Judd, A. Oliva, Torralba, and F. Durand, "What do different evaluation metrics tell us about saliency models," *arXiv preprint arXiv:1604.03605*, 2016.

[9] J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 6, pp. 679-698, 1986.

[10] S. Carpendale, "Evaluating Information Visualizations" *Information Visualization: Human-Centered Issues and Perspectives*, A. Kerren, J. Stasko, J.-D. Fekete, C. North (Eds.), Springer, pp. 19-45, 2008.

[11] H. Chen, S. S. Tsai, G. Schroth, D. M. Chen, R. Grzeszczuk, and B. Girod, "Robust text detection in natural images with edge-enhanced maximally stable extremal regions." *18th IEEE International Conference on Image Processing (ICIP),* pp. 2609-2612. IEEE, September 2011.

[12] C. E. Connor, H. E. Egeth, and S. Yantis, "Visual attention: Bottom-up versus top-down," *Current Biology*, vol. 14, pp. R850-R852, 2004.

[13] R. Etemadpour, B. Olk, and L. Linsen, "Eye-tracking investigation during visual analysis of projected multidimensional data with 2D scatterplots," *International Conference on Information Visualization Theory and Applications (IVAPP),* pp. 233-246. IEEE, 2014.

[14] M. D. Fairchild and R. S. Berns, "Image color-appearance specification through extension of CIELAB" Color Research & Application, vol. 18, pp. 178-190, 1993.

[15] J. H. Goldberg and J. Helfman, "Comparing information graphics: A critical look at eye tracking," *Proceedings of the 3rd BELIV'10 Workshop: BEyond time and errors: novel evaLuation methods for Information Visualization,* pp. 71-78. ACM, April 2010.

[16] J. H. Goldberg and J. Helfman, "Eye tracking for visualization evaluation: Reading values on linear versus radial graphs," *Information visualization*, vol. 10, 182-195. 2011.

[17] A. Gonzalez, L. M. Bergasa, J. J. Yebes, and S. Bronte, "Text location in complex images," *21st International Conference on Pattern Recognition (ICPR),* pp. 617-620. IEEE, November 2012.

[18] M. J. Haass, A. T. Wilson, L. E. Matzen and K. M.& Divis, "Modeling Human Comprehension of Data Visualizations," *International Conference on Virtual, Augmented and Mixed Reality*, pp. 125-134. Springer International Publishing, July 2016.

[19] K. W. Hall, C. Perin, P. G. Kusalik, C. Gutwin, and S. Carpendale, "Formalizing emphasis in information visualization," *Computer Graphics Forum*, vol. 35, pp. 717-737. 2016.

[20] J. Harel. "A saliency implantation in MATLAB" http://www.vision.caltech.edu/~harel/share/gbvs.php. 2017.

[21] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," *Proceedings of Neural Information Processing Systems (NIPS),* pp. 545-552, 2006.

[22] J. M. Henderson, J. R. Brockmole, M.S. Castelhano, and M. Mack, "Visual saliency does not account for eye movements during visual search in real-world scenes" *Eye Movements: A Window on Mind and Brain*, pp.537-562, 2007.

[23] H. Ho, I. Yeh, Y. Lai, W. Lin, and F. Cherng, "Evaluating 2D flow visualization using eye tracking," *Computer Graphics Forum*, vol. 34, pp. 501-510. 2015

[24] L Itti and C. Koch, "Computational modelling of visual attention," *Nature Reviews Neuroscience, vol. 2*, pp. 194-203, 2001.

[25] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis" *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. *11*, pp. 1254-1259, 1998.

[26] H. Jänicke and M. Chen, "A salience-based quality metric for visualization," *Computer Graphics Forum*, vol. 29, pp. 1183-1192. Blackwell Publishing Ltd., 2010.

[27] T. Judd, F. Durand, and A. Torralba, "A benchmark of computational models of saliency to predict human fixations" MIT Technical Report, 2012.

[28] D. Kahneman and D. Chajczyk, "Tests of the automaticity of reading: Dilution of Stroop effects by color-irrelevant stimuli," *Journal of Experimental Psychology: Human Perception and Performance,* vol. 9, pp. 497, 1983.

[29] K. Kurzhals, B. Fisher, M. Burch, and D. Weiskopf, "Evaluating visual analytics with eye tracking," *Proceedings of the Fifth Workshop on Beyond Time and Errors: Novel Evaluation Methods for Visualization*, pp. 61-69. ACM, 2014.

[30] H. Lam, E. Bertini, P. Isenberg, C. Plaisant, and S. Carpendale, "Empirical Studies in Information Visualization: Seven Scenarios" *IEEE Trans. Visualization and Computer Graphics*, vol. 18*,* pp. 1520– 1536, 2012.

[31] G. E. Legge, S. J. Ahn, T.S. Klitz, and A. Luebker, "Psychophysics of reading—XVI. The visual span in normal and low vision" *Vision Research,* vol. 37*,* pp. 1999-2010, 1997.

[32] Y. Li and H. Lu, "Scene text detection via stroke width," *21st International Conference on Pattern Recognition (ICPR),* pp. 681-684. IEEE, November 2012.

[33] G. D. Logan, "Automaticity and reading: Perspectives from the instance theory of automatization" *Reading & Writing Quarterly: Overcoming Learning Difficulties,* vol. 13, pp. 123-146, 1997.

[34] S. Lu, T. Chen, S. Tian, J. H Lim and C. L Tan "Scene text extraction based on edges and support vector regression" *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 18, pp. 125-135, 2015.

[35] C. M MacLeod, "Half a century of research on the Stroop effect: An integrative review" *Psychological bulletin,* vol. 109, p. 163, 1991.

[36] J. Matas, O. Chum, M. Urban & T. Pajdla, "Robust wide-baseline stereo from maximally stable extremal regions," *Image and Vision Computing*, vol. 22, pp. 761-767, 2004.

[37] L. E. Matzen, M. J. Haass, K. M. Divis and M.C. Stites "Patterns of attention: How data visualizations are read," *Augmented Cognition. Enhancing Cognition and Behavior in Complex Human Environments*, D. D. Schmorrow and C. M. Fidopiastis, eds., pp. 176-191. Springer, 2017.

[38] L. E Matzen, M. J. Haass, J. Tran, and L. A. McNamara, "Using eye tracking metrics and visual saliency maps to assess image utility," *Proc. Human Vision and Electronic Imaging (HVEI) XXI*, 2016.

[39] E. N. Merieb and K. Hoehn, "Human Anatomy & Physiology 7th Edition," Pearson International Edition, 2007.

[40] L. Neumann and J. Matas, "Real-time scene text localization and recognition," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 3538-3545. IEEE, June 2012.

[41] T. Ogawa and H. Komatsu, "Target selection in area V4 during a multidimensional visual search task" *Journal of Neuroscience*, vol. 24, pp. 6371- 6382, 2004.

[42] N. Pinto and D.D. Cox, "Beyond simple features: A large-scale feature search approach to unconstrained face recognition" *IEEE Automatic Face and Gesture Recognition*, 2011.

[43] Y. Pinto, A. R. van der Leij, I.G. Sligte, V. A. F. Lamme, and H. S. Scholte, "Bottom-up and top-down attention are independent," *Journal of Vision,* vol. 13, pp. 1-14, 2013.

[44] B. Strobel, S. Saß, M. A. Lindner, and O. Köller, "Do graph readers prefer the graph type most suited to a given task? Insights from eye tracking," *Journal of Eye Movement Research*, vol. 9, pp. 1-15. 2016.

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

[46] Ware, Colin. *Information visualization: perception for design*. Elsevier, 2012.

[47] A. Yarbus, *Eye Movements and Vision*. New York City: Plenum Press, 1967.

[48] J. Zhang and S. Sclaroff "Exploiting surroundedness for saliency detection: A boolean map approach," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2015.

[49] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, "Sun: A Bayesian framework for saliency using natural statistics," *Journal of Vision,* vol. 8, pp. 32-32, 2008.