Measuring and Modeling the Feature Detection Threshold Functions of Colormaps

Colin Ware, Terece L. Turton, Roxana Bujack, Francesca Samsel, Piyush Shrivastava, and David H. Rogers

Abstract—Pseudocoloring is one of the most common techniques used in scientific visualization. To apply pseudocoloring to a scalar field, the field value at each point is represented using one of a sequence of colors (called a colormap). One of the principles applied in generating colormaps is uniformity and previously the main method for determining uniformity has been the application of uniform color spaces. In this paper we present a new method for evaluating the feature detection threshold function across a colormap. The method is used in crowdsourced studies for the direct evaluation of nine colormaps for three feature sizes. The results are used to test the hypothesis that a uniform color space (CIELAB) will accurately model colormapped feature detection thresholds compared to a model where the chromaticity components have reduced weights. The hypothesis that feature detection can be predicted solely on the basis of luminance is also tested. The results reject both hypotheses and we demonstrate how reduced weights on the green-red and blue-yellow terms of the CIELAB color space creates a more accurate model when the task is the detection of smaller features in colormapped data. Both the method itself and modified CIELAB can be used in colormap design and evaluation.

Index Terms—Colormapping, color perception

1 Introduction

One of the most common and effective methods for visualizing scientific data is using a color sequence, commonly called a *colormap*, to encode scalar values in univariate map data [40], [53], [62]. A set of examples in Fig. 1 shows the same sea surface height data rendered using nine different colormaps. What makes a good colormap? Clearly, to some extent this depends on the way it will be used. Three broad task categories can be identified:

- *Pattern Perception:* The first and broadest task category is feature or pattern perception [4], [58]. The patterns that may be of scientific interest are essentially infinite. A researcher may be interested in feature shapes, and how large or small they are in terms of spatial size or amplitude. Basic to all pattern perception is the feature detection threshold—if the features making up a pattern cannot be seen, the pattern cannot be seen.
- *Value reading tasks and value localization tasks:* The value reading task is to determine the data value at a point on a map, usually by means of a key [56], [58]. An example of such is a weather map with color-coded temperatures; observing the color of a point on the map and visually matching that color to a color key allows the temperature to be estimated. The value localization task is the reverse of this [34]. A value is given and the task is to find locations on the map corresponding to that value.
- *Categorization task:* Sometimes colors are used to visually categorize data [9], [50]. For example, in large-scale geographical maps, greens roughly characterize low plains whereas browns characterize mountainous regions. Shades of blue are used for ocean depths.

Much of the work that has been done on the design of colormaps has focused on design principles, not on tasks per se. These principles include *order*, *smoothness*, *uniformity*, and *discriminative power* [11], [55]. Order in a global sense is the degree to which a sequence is perceived as progressing through colors in a particular direction [4], [38], [61]. The options available in the ColorBrewer palettes for cartography are good examples of ordered palettes [9]. The widely used (and equally widely criticized) rainbow colormap has no overall perceptual order [8], [44], [58]. Smoothness refers to the extent to which a colormap has no distinct boundaries in the sequence of colors [30], [53], [54]. Uniformity refers to the extent colors equally separated on the colormap correspond to equal perceptual distances [8], [21], [29], [30], [41]. Discriminative power refers to how many perceivably different colors are traversed by the colormap. Usually this is defined as the number of just noticeable differences (JND) over the entire sequence [11].

In the present study, we are concerned with the discriminative power function of a colormap (how well it resolves

- *C. Ware and P. Shrivastava are with the University of New Hampshire, Durham, NH 03824. E-mail: cware@ccom.unh.com, piyush.shrivastava9@gmail.com.*
- *T.L. Turton, R. Bujack and D.H. Rogers are with Los Alamos National Laboratory, Los Alamos, NM 87545. E-mail: {tlturton, bujack, dhr}@lanl.gov.*
- *F. Samsel is with Center for Agile Technology, University of Texas at Austin, Austin, TX 78712, USA. E-mail: figs@cat.utexas.edu.*

Manuscript received 7 Aug. 2017; revised 6 June 2018; accepted 3 July 2018. Date of publication 19 July 2018; date of current version 31 July 2019. (Corresponding author: Colin Ware.) Recommended for acceptance by V. Interrante. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below. Digital Object Identifier no. 10.1109/TVCG.2018.2855742

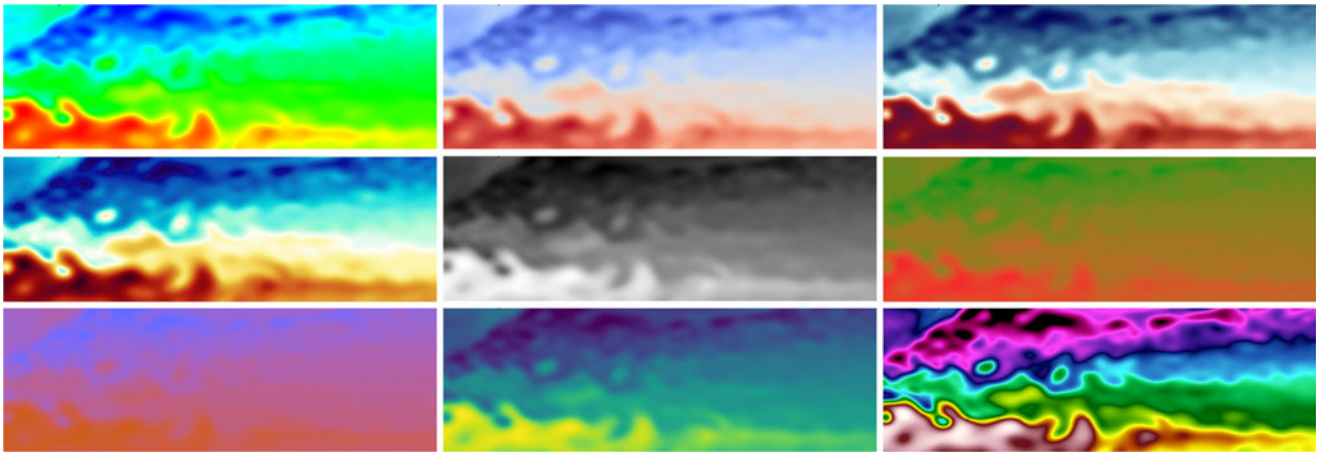

Fig. 1. Sea surface height rendered in the nine test colormaps in this paper. Using the acronyms introduced in Section 4.2, the data is rendered in: (top row, left to right) RA, CW, ECW, (middle row, left to right) BOD, GP, GR, (bottom row, left to right) BY, VI, TH.

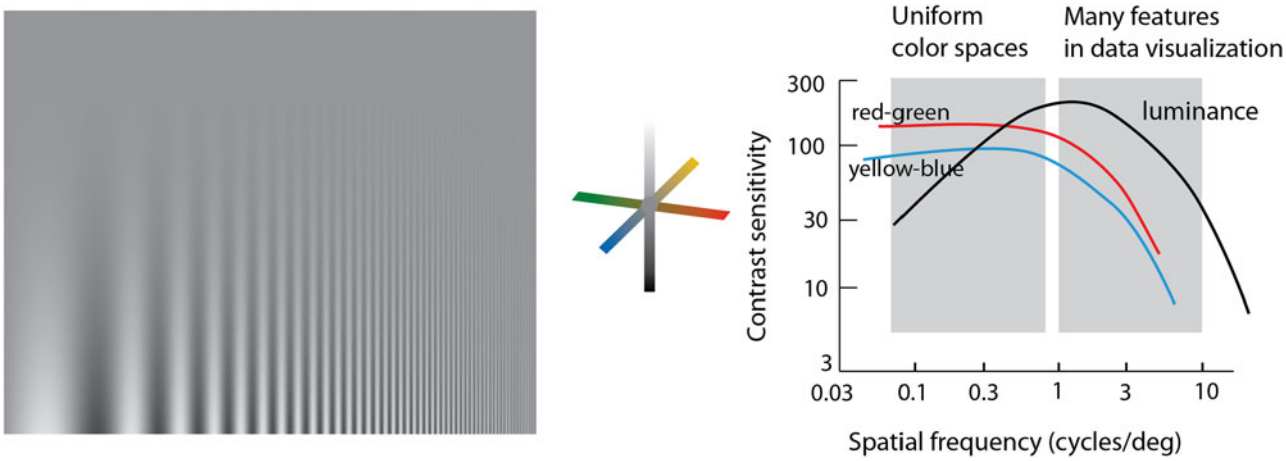

Fig. 2. Left: A pattern illustrating how human spatial patterns sensitivity falls off for both low and high spatial frequencies. Right: human pattern sensitivity for the different color channels as a function of spatial frequency (Adapted from Mullen [32]). Uniform color spaces were based on measurements with large stimuli containing low spatial frequencies. Many visualization are dominated by higher spatial frequencies.

features across its extent). Local discriminative power is analgous to contrast sensitivity, a term commonly used in psychophysics to denote the inverse of a contrast threshold (the minimum contrast that can be resolved by a human subject). We also derive a metric of overall discriminative power—the overall capacity of a colormap to resolve features.

The contribution of this paper is a straightforward method for evaluating the discriminative power function of a colormap across its extent. This is applied for a range of feature sizes and the results used to develop a modified version of CIELAB that more accurately models human perception of features in colormapped data. We also show how modified CIELAB can be applied in a colormap design tool.

2 Background and Related Work

The most common method that has been advocated for creating perceptually uniform color sequences has been the application of uniform color spaces (UCSs) such as CIELUV, CIELAB, CIEDE2000 and CIECAM [10], [15], [24], [25], [27], [31]. The most popular uniform color spaces are mathematical transformations of the CIE XYZ coordinates, constructed such that metric differences between pairs of colors in the space correspond to experimentally determined perceptual differences from user studies. A uniform colormap defined using a UCS is a sequence where adjacent colors have equal separations in that UCS. The overall discriminative power of the colormap defined by a UCS is simply the total length of the path of the colormap in the UCS [11], [21], [35], [40].

Although using a UCS to understand the power of colormaps to resolve features is common, there are reasons for thinking that the standard UCSs will not actually provide a good basis for measuring or creating uniformity for many of the features that are critical in scientific data. Uniform color spaces were based on measurements between two quite large patches of uniform colors [26], [33], [39]. They were intended for the paint and fabric industries and there is reason to believe that they will not, in fact, provide a good basis for assessing the quality of colormaps in terms of either their uniformity, or their ability to allow people to resolve small features in data. The problem has to do with the dependence of feature discriminative power of different color channels, as can be seen in Fig. 2. Opponent color theory holds that human color vision can be characterized in terms of three color-opponent channels, the luminance (black-white) channel, the green-red channel, and the yellow-blue channel [16]. Various measurements have shown that the different channels have very different characteristics. For example, the

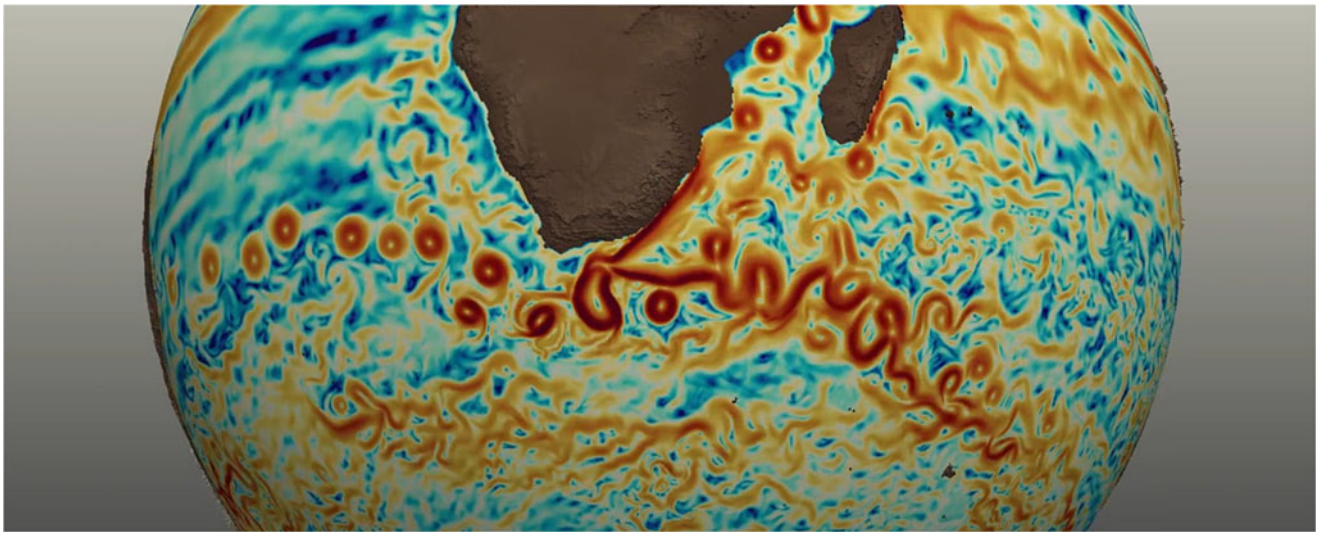

Fig. 3. A colormapped image showing current speed in the southern oceans. The important features such as eddies and jets are quite small.

luminance channel is better at revealing shape-from-shading information and patterns in motion [3], [14]. Most relevant here are findings that the different channels have very different feature discriminative powers, particularly at higher spatial frequencies [32], [36].

Vision researchers have been characterizing the human visual system in terms of its ability to resolve sinusoidal patterns of different spatial frequencies since the late 1960s [6], [12]. The left-hand image of Fig. 2 illustrates this concept. Campbell and Robson [12] showed that human ability to resolve complex patterns is solely determined by sensitivity of the constituent frequencies. For luminance modulated patterns there is both a distinct high frequency and low frequency fall off in sensitivity. Fig. 2 shows (on the right) the results from a study that investigated the feature discriminative power of the green-red and yellow-blue channels in addition to the luminance channel [32]. Both this and a study by Poirson and Wandell [36] show that at spatial frequencies of one cycle/degree and above, the luminance channel has far greater sensitivity to patterns in comparison with the color channels. Since it is the case that the UCS measurements were made with larger patterns (2 and 10 deg [33]) containing low spatial frequencies this would suggest that using UCSs to design colormaps will considerably overweight the contributions of the green-red and yellow-blue channels to both uniformity and overall discriminative power. This means that they may be unsuitable as a basis for designing colormaps for data such as the ocean eddies and currents illustrated in Fig. 3.

The differences between luminance and chromatic channels have been noted by researchers who have suggested that high spatial frequency patterns should be represented mainly by luminance variation [40], [43], [44] and low spatial frequencies are better expressed by chromatic variation [4], [40]. Others, from Stevens on, have noted that fewer steps can be resolved in chroma than can be resolved in luminance [4], [5], [43], [49]. *Chroma* is a technical term that refers to the vividness of a color—its distance from a neutral gray of the same luminance. Saturation is often informally used to refer to the same quality but technically has a somewhat different definition. Also, various researchers have proposed that aside from the issue of resolution, luminance variation is more suitable for form perception [40], [42], [58].

The two studies that bear most directly on the detection of patterns in colormapped data are by Rogowitz et al. [43] and Kalvin et al. [17]. Rogowitz et al. measured detection thresholds using large Gaussian patterns placed 3 deg from the center of fixation for a number of colormaps. The Fourier transform of a Gaussian is also a Gaussian and the patterns used would have had dominant spatial frequency components below 0.5 cycles/deg. In addition to testing a rainbow colormap they compared the influence of linear changes in hue versus chroma versus value in Munsell and CIELAB space. They found that hue based colormaps such as the rainbow perform worse with respect to uniformity than the luminance and chroma based ones. Kalvin et al. used Gabor stimuli at 0.2 and 4 cycles/deg. For high spatial frequency patterns they found an increase in threshold for saturation and hue variation. The results for luminance variation, however, were puzzling. For grey scales defined by the HSV model there was little increase in threshold, whereas when the grey scale was defined by CIELAB there was a substantial increase in threshold. They offer no explanation for this striking discrepancy. We build on this work, although with a very different methodology, using sinusoidal patterns.

Other evidence for the importance of luminance in form perception comes from a study by Ware [58] in which study participants were presented with a variety of patterns (parabola, saddle, ridge, etc) and asked to rate how well the underlying shapes were represented using Likert scales. Overall, the grayscale colormap was judged the most effective at representing underlying shape and the author argued that this was because the luminance channel is most relevant to form perception. However, this was a subjective and not an objective measurement. Rogowitz and Kalvin [42] encoded a photograph of a human face with various colormaps in order to evaluate them. Faces colormapped with monotonically increasing luminance were judged to appear most natural. Faces are very special patterns for which the brain has a dedicated processing area (e.g., [18], [28]), but it is not clear if judgments of naturalness in faces generalize to the more abstract patterns within scientific

visualization. Kindlmann et al. [19] used the human facial recognition ability to generate a method for personalized luminance matching on uncalibrated displays. It can be used for generating colormaps with predefined properties w.r.t. luminance, for example, monotonicity or constancy.

Related papers by Stone, Szafir, and collaborators [51], [52] emphasized the influence of symbol size on the discrimination of colored symbols for applications in the design of discrete symbols used in information visualization. They adapted CIELAB to match their results to provide an "engineering model" for use in the design of colored symbols. Our modeling approach is similar.

Color-plane variation is often broken down into hue and chroma components. Hue denotes the cycle of colors, from red to yellow to green to blue to purple and back to red. Chroma denotes the vividness of colors, or how much they differ from neutral grays of the same luminance. It has often been noted that the commonly used spectrum approximation, as a whole is not perceptually ordered, although parts of it are [8], [58]. Color channels can also carry perceptually ordered information in the form of chroma, as can double ended colormaps (e.g., red-green or yellow-blue) [4], [5], [17], [43]. It is also the case that most of the colormaps in use contain variation in both lightness and chromaticity variation over their extent. (Note: Chromatiticity refers to any non luminance variation in color). Few people in scientific visualization use a simple grayscale colormap. In the present paper, we are concerned solely with the ability of colormaps to enable feature detection.

Finally, it is worth considering that by stretching and compressing a colormap at different points it will always be possible to take any colormap and make it uniform, whereas the overall discriminative power should not dramatically change, at least according to some metrics [11]. It is also important to note that in many cases, colormaps are deliberately made to be non-uniform in order to emphasize a particular value range in the data, for example, as in [7], [46]. Nevertheless, there are good reasons for designing uniform sequences as most people use colormaps unmodified, lacking the tools to selectively stretch and compress parts of the sequence to suit their needs.

3 Contribution and Hypotheses

The work presented here had a number of goals:

1) Develop a method for directly measuring the feature detection threshold functions of colormaps that can be used for differently sized features. Our method is designed to build a bridge between spatial perception theory and colormap evaluation.
2) Conduct a study evaluating a set of colormaps for differently sized features.
3) Determine if a modified UCS can model the results.
4) Show how the method can be applied in a colormap design tool.

Based on the differing spatial sensitivities of color opponent channels compared to the luminance channel [32], [36], we can make specific predictions relating to the modifications needed for USCs to accurately model the results. Where the task is to identify small features in colormapped data, UCSs will fail to accurately model feature detection thresholds. Specifically they will overweight the contributions of the green-red and blue-yellow channels. We test this hypothesis by modeling the data using CIELAB with modified weights on the a^* and b^* terms. A corollary of this hypothesis is that colormaps with the greatest variation in luminance will have the most overall feature discriminative power. We also test the hypothesis that a luminance only model (L^*) can account for the data.

A preliminary report on the method appeared in Ware et al. [59]. That short paper introduced the basic method and applied it to seven colormaps. Here we provide a much more complete account. We have extended the method, applying it to additional colormaps designed to specifically test in the green-red and yellow-blue directions. In addition, this paper applies the method to multiple feature sizes to test for size dependency effects. All of the UCS modeling work is presented here for the first time.

4 Method

The basic study method was briefly introduced in Ware et al. [59]. The following description is more complete, describing the additional feature sizes and additional colormaps.

4.1 Test Patterns

The method is based on the test pattern illustrated in Fig. 4. This has columns of features that reduce in contrast from bottom to top. The point at the top of any column where the pattern becomes invisible represents the local discriminative power. Notice how the patterns fade out at approximately the same height for the gray colormap, but at very different heights for the rainbow colormap. Sets of these patterns are used to estimate feature detection thresholds at 30 points along the colormap.

The basic test pattern is an artificial data image with the following properties. The background of the image is a linear ramp, increasing from 0.1 to 0.9 from left to right. Note that the reason this does not range from zero to one is to avoid truncation of the target patterns. Added to the background ramp is a set of six equally spaced columns of features. These contain oblique sinusoidal patterns as shown. For each column, contrast increases according to a power law from top to bottom:

$$a = c * 2^{(1+(y-s)/p)}, \tag{1}$$

where c is the starting amplitude, y is the distance from the top of the image, s is the position at which the pattern starts, and p is the amplitude doubling interval. All units are in pixels. For our 600×600 test patterns, $c = 0.001$, $s = 40$ pixels, and $p = 80$ pixels. Since the pattern varied over 560 pixels, this yielded seven doublings, or a factor of 128 from top to bottom. A value, v, between 0 and 1 is computed at each pixel using

$$v = r + 0.5ag(sin(2\sqrt{2}\pi(x+y))\lambda), \tag{2}$$

where r is the ramp value, x and y are pixels, λ is the spatial wavelength and g is a Gaussian distribution:

$$g = e^{-[3((x-x_0))(2\lambda)]^2}. \tag{3}$$

Here x_0 represents the horizontal position of a particular feature column.

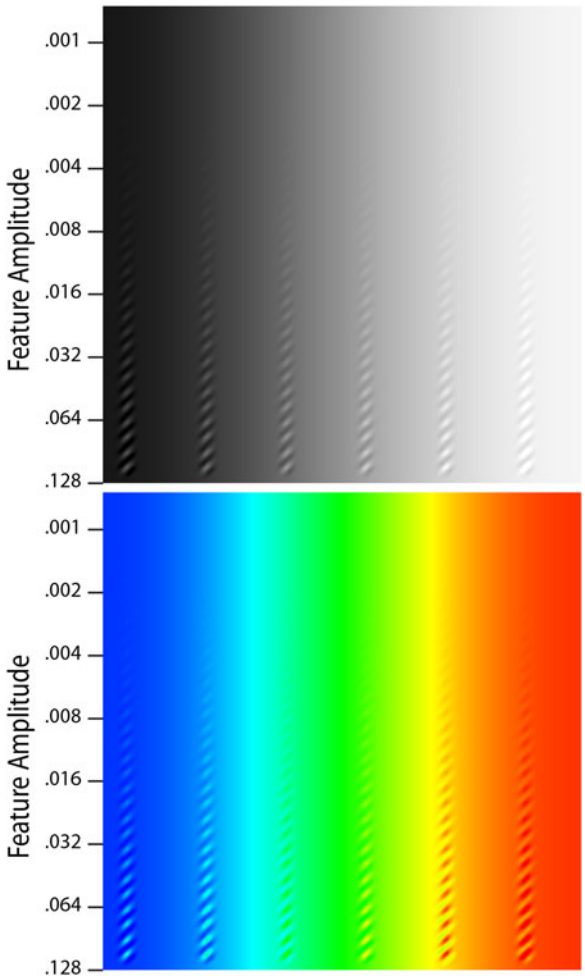

Fig. 4. The test pattern we use shown with a gray colormap (top) and a rainbow colormap (bottom) for the 15 pixel feature size. Six vertical columns of sinusoidal features are shown. In each column the contrast decreases by a factor of two, every 80 pixels. Clicking on the point where a feature column becomes invisible yields a measure of the detection threshold for features having the spatial frequency of the pattern. A set of five images for each colormap yields 30 sample points across the colormap.

There were three pattern wavelengths used in this study: 10, 15, and 45 pixels. In each test pattern, for the 10 and 15 pixel data, six discrete vertical stripes of the sine pattern were constructed as illustrated in Fig. 4 for the 15 pixel data. Stripes were horizontally separated by 100 pixels. Sets of 5 such patterns were generated for each colormap with starting offsets of 10, 20, 50, 70, and 90 pixels. Because the patterns were considerably wider for the 45 pixel feature set, we replaced each single test pattern with two, each 600 pixels wide, testing the lower and upper ranges of each colormap respectively as shown in Fig. 5. In these, the vertical feature stripes were separated by 200 pixels, and the column offsets were adjusted appropriately. Initially, a spatial frequency of 15 pixels was studied. The 45 pixel and 10 pixel values were chosen to extend the range of spatial frequencies investigated.

To render images used as stimuli, each colormap table was expanded in software to a 1000 entry RGB look up table, using linear interpolation. The color of each pixel in a stimulus image was obtained by multiplying the data value at that point by 1000 and using the result to index into the table and obtain the corresponding RGB value. For transformations to

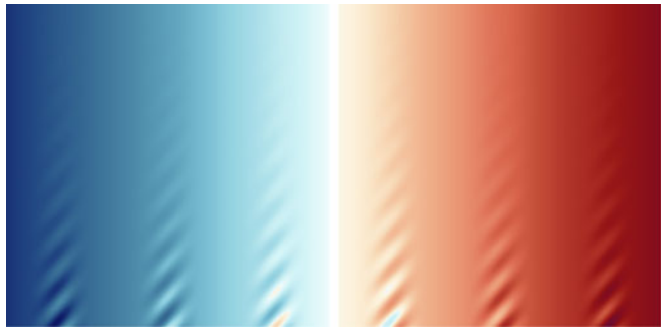

Fig. 5. An example of the two test patterns that were combined to test the 45 pixel feature set, shown in the CW colormap. Although the size has been reduced here to conserve space, each was shown at the same size as the ones in Fig. 4.

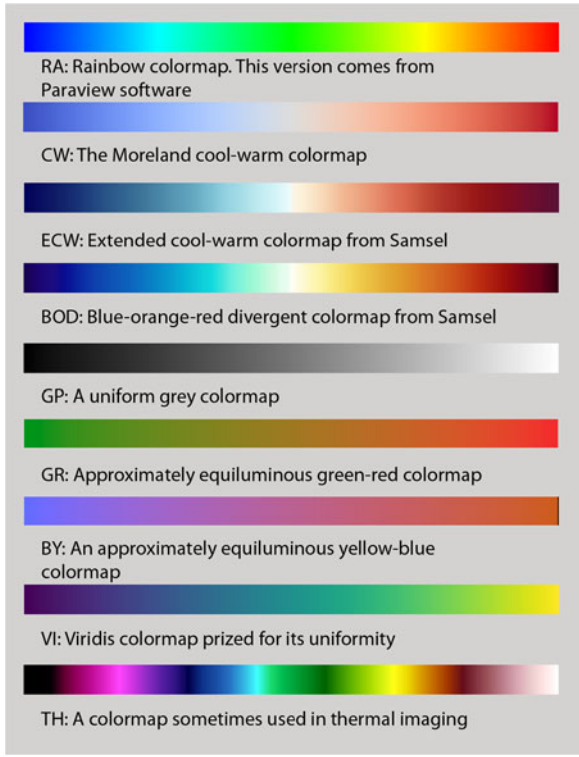

Fig. 6. The nine color sequences used in the study (from top to bottom): RA:rainbow, CW:cool/warm, ECW:extended cool/warm, BOD:blue/orange divergent, GP:grayscale, GR:green-red, BY:yellow-blue through red/blue, VI:Viridis, TH:thermal.

CIE XYZ and CIELAB it was assumed that all screens met the sRGB standard and that the reference white (for CIELAB) was defined by the maximum values on R,G and B respectively.

4.2 Colormaps

Nine colormaps, shown in Fig. 6, were used in the current study with three different spatial frequencies of the test pattern. These colormaps were chosen for a variety of reasons.

- The rainbow (RA), one of many spectrum-based versions, is from ParaView [1].
- The Moreland cool/warm (CW) [30] is a commonly used example of a double-ended colormap.
- The extended cool/warm (ECW) and blue/orange divergent (BOD) sequences by artist F. Samsel and the Data Science at Scale Team (DSS) at LANL [47]

are more recent examples of double-ended colormaps designed to maximize feature resolution. Both of these have a large luminance variation.

- The grayscale (GP) was constructed to have equal steps in CIELAB L* ranging from 0-100. The values were converted to XYZ and then to RGB assuming the sRGB monitor standard.
- The green-red (GR) was designed to vary only on the green-red color channel. It has equal steps in CIELAB a* (the representation of the green-red channel). It varies from -60.5 to 73.0. L* and a* are constant at 53 and 50 respectively.
- The blue-red (BY) colormap. The reason for choosing blue-red, rather than yellow-blue, is because of the shape of the gamut of R,G,B colors in CIELAB, with the maximum range occurring between red and blue. However, this colormap only varies in b*, the CIELAB representation of the blue-yellow channel. It has equal steps in b* from -78 to 55. L* and a* are constant at 53 and 42 respectively. We have labeled it BY in order to emphasize that it varies in the blue-yellow direction defined by CIELAB.
- The Viridis (VI) [57] colormap is an example of a widely used uniform colormap designed to cycle through a number of hues.
- The Thermal (TH) sequence is sometimes used in Thermal imaging. It has the unique property of traversing most of the luminance range seven times, which, according to the hypothesis in Section 3, should give it extreme overall detection power.

The colormaps used are given as [V, R, G, B] tables in supplementary material, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TVCG.2018.2855742 where V varies between 0 and 1.

Because of the very low discriminative power of the yellow-blue sequence we doubled the starting contrast to 0.002 to keep it within the range of the test pattern. Similarly, because of the very high threshold discriminative power of the Thermal imaging sequence we halved the starting contrast to 0.0005.

4.3 Task

For each test pattern, the participant's task was to click on each of the six points in the columns where the vertical pattern became invisible.

4.4 Participants

Participants were solicited on Mechanical Turk [2] and paid $0.85. In total, 560 unique participants were collected across the three different feature datasets with 55.9 percent male, 43.2 percent female, and 0.9 percent unspecified participants. Participants ranged from 18 to 73 years of age with a mean age of 36. Since this study involved color, precautions were taken to minimize any potential contamination due to color vision deficiencies (CVD). A fuller discussion and validation of those precautions can be found in Section 7.1.

4.5 User Study Procedure

The experimental procedure closely followed the method laid out in [59]. The study itself was coded using the heat map question in Qualtrics survey software [37] and the studies were launched on Mechanical Turk using the TurkPrime interface [22]. Using built-in Qualtrics functionality, participants on mobile devices were blocked from taking the study. Participants were asked to check that the browser was on 100 percent zoom and to place themselves 50 cm from the screen (with advice for how far that was for an average male or female). Only 5.5 percent of participants had a screen resolution of 1280×720 or under. The most common screen resolutions were 1360×768 (38.9 percent), 1920×1080 (28.8 percent), and 1600×900 (10.4 percent). The average screen resolution was 1580×920. For typical laptop and desktop screens this yields the cycles/degree values given in Table 1.

TABLE 1
Estimated Cycles/Degree for the Three Features Sizes

Feature Size:	10 pixel	15 pixel	45 pixel
Typical laptop	5.11	3.41	1.14
Typical desktop	3.87	2.58	0.86

Each colormap was tested at 30 data points as discussed above, with five stimuli images in the case of the 10 pixel and 15 pixel data, and 10 stimuli images in the case of the 45 pixel data. The test patterns were shown one at at time, with a set of five/ten testing a single colormap given sequentially. The individual stimuli images were presented in randomized order. Each participant saw all stimuli for one to three randomly chosen colormaps. A participant was allowed to take the study again for a different feature size. The number of participants per colormap ranged from 21 to 35 participants.

Data was manually scanned to remove participants whose click pattern indicated they either did not understand the task (click once per column) or were not faithfully completing the task (e.g., always clicked at the top/bottom of the columns). Additionally, participant clicks were required to be no more than twice the feature size away from the nominal horizontal center of each column. These validation checks removed 9 percent of the participants.

5 Results

The results are summarized in Fig. 7. These plots show the log discriminative power, averaged across participants for each of the nine colormaps and for each of the three pattern sizes. We are using a log scale for these plots and for most of the analysis, both because the stimulus test pattern contrast was exponentially scaled on the vertical axis, and because it better expresses the range of variation both within and across colormaps. Discriminative power is the inverse of the measured threshold (the minimal amplitude of the sine pattern that can be resolved), and this is analogous to contrast sensitivity used in psychophysics. It is also analogous to $\Delta E/\Delta s$, where ΔE is the measure of difference between two colors in a uniform color space, and Δs is the distance along a colormap normalized to a length of 1.0. The data are somewhat arbitrarily separated into two groups: patterns that vary monotonically in luminance, and those that do not. The other reason for this separation is having all the data on a single plot was overly cluttered.

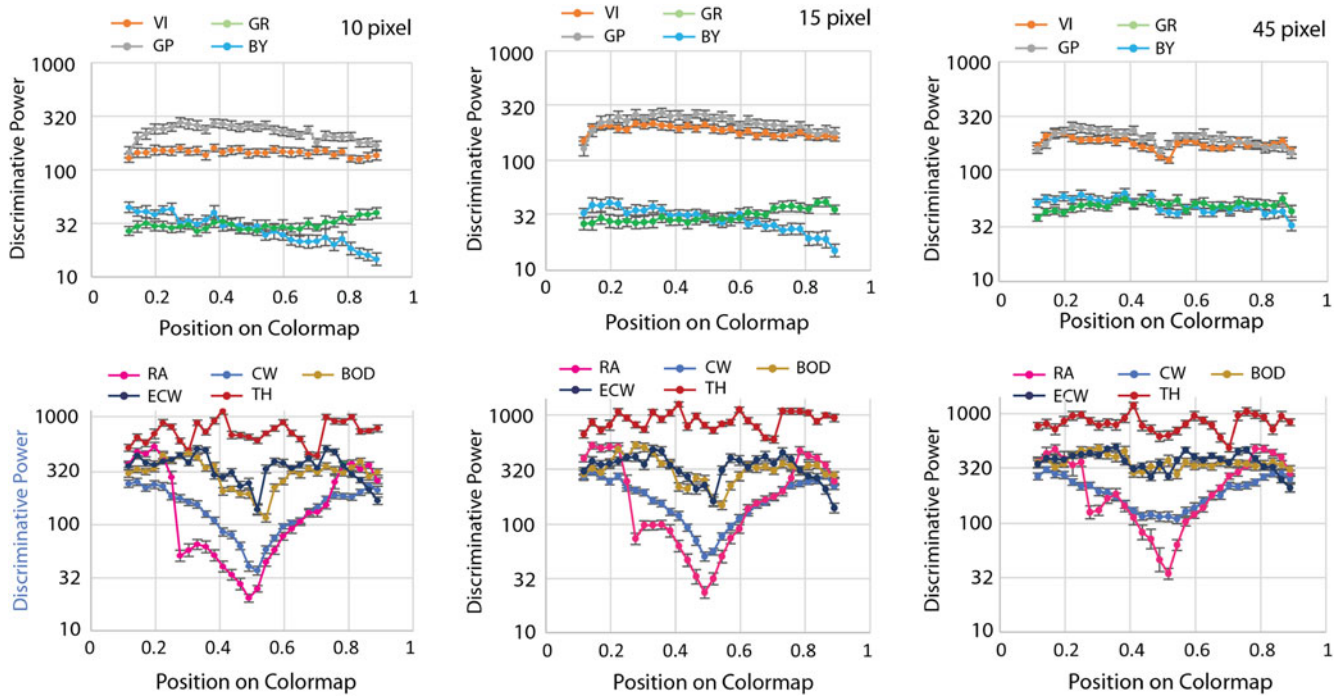

Fig. 7. The disciminative power functions for the (left to right) 10, 15, and 45 pixel data patterns.

Inspection of these empirical functions tells us a great deal about the different colormaps we tested. First, we observe some of the characteristics of the curves' shapes that are common across the three feature sizes.

- The Rainbow colormap (RA) is extraordinarily non-uniform; in its middle section it has 1/16th of the feature discriminative power that it does at either end.
- As expected, the Thermal colormap has the greatest overall discriminative power, not surprising as it has the greatest path length along the luminance direction. It goes from dark to light and back again several times.
- The Samsel divergent colormap (BOD) has the next greatest discriminative power, except in the middle sections where it worse than the gray scale. Divergent ECW also has high discriminative power except in the middle and at the high end.
- The Viridis (VI) colormap is the most uniform of all those tested.
- The green-red (GR) and blue-red (BY) colormaps have very low discriminative power. The BY curve shows reduced discriminative power at the red end for smaller features.
- The gray colormap (GP) is somewhat less uniform compared to Viridis, but it also has greater discriminative power over most of its length.
- The Moreland cool-warm (CW) colormap is not uniform for features of this size, even though it was designed to be uniform. The curve is somewhat flatter for the large (45 pixel) feature sizes.

One feature of note in the 45 pixel data is the notch that appears in both the gray and viridis colormaps at the center. We believe that this is an unfortunate artifact arising because the colormaps were divided into two parts for the 45 pixel feature size test.

Contrast sensitivity is the reciprocal of the contrast threshold. It is a measure of the discriminative power of a color map and can be equated to ΔE values in uniform color space. To allow us to compare the colormaps, average discriminative power was computed on a subject-by-subject basis for the nine colormaps at the three frequencies tested. This is analogous to the overall trajectory length of a colormap in a uniform color space. One caveat is that our method only evaluates the middle 80 percent of a colormap. These averaged results are shown in Fig. 8. As predicted, the thermal colormap has by far the greatest average discriminative power, followed by the two Samsel divergent colormaps (ECW and BOD). The green-red and yellow-blue colormaps have very low average discriminative power, also as predicted.

To statistically compare the different colormaps in terms of their average discriminative power, we ran a 2-way ANOVA (feature size, colormap) on the contrast sensitivity

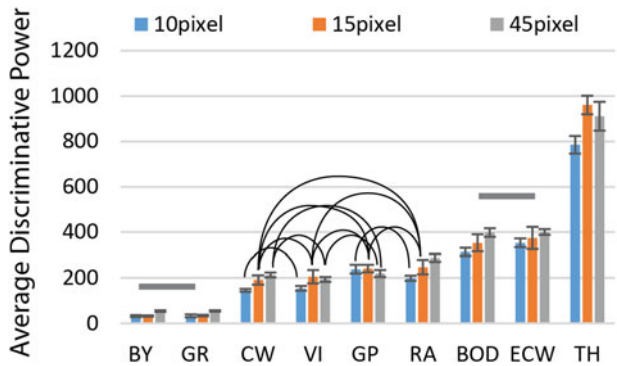

Fig. 8. The mean discriminative power for the nine colormaps tested at the three spatial frequencies. Error bars correspond to 95 percent CIs. The solid bars indicate where the Tukey HSD found no significant differences between those mean discriminative power for each of the three feature size. The arcs indicate where colormaps were not significantly different for specific feature sizes.

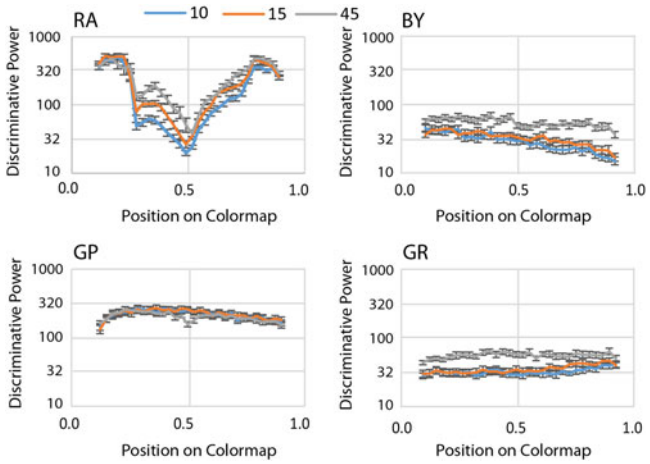

Fig. 9. The detection threshold curves for the rainbow, grayscale, the yellow-blue and green-red colormaps. Note how the thresholds drop as a function of feature size for the green-red and yellow-blue colormaps and for parts of the rainbow.

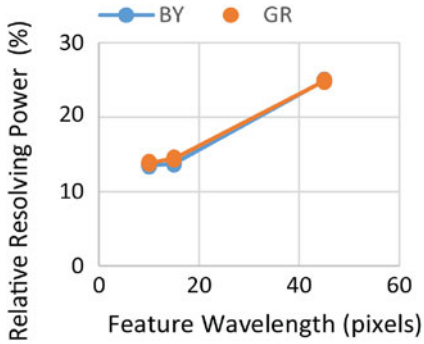

Fig. 10. Relative resolving power across all three feature sizes. The ratio of the average contrast sensitivities (with respect to the gray) is plotted for the BY and GR colormaps.

results. Both main effects and the interaction were highly significant. For feature size ($F(2,703) = 35.6$, $p < 0.001$); for colormap ($F(8,703) = 1155$, $p < 0.001$); for the interaction $F(16,703) = 8.59$, $p < 0.001$). We also ran Tukey HSD tests for the differences between colormaps separately for each of the three feature sizes. The results of the HSD tests are indicated by grouping lines shown above the bars in Fig. 8. A solid bar indicates that the colormaps are not different for all three of the features sizes. The arcs show where colormaps failed to differ for a specific feature size. Overall there are four groupings. The BY and GR colormaps have lower average discriminative power than any of the others. The thermal colormap has the greatest average discriminative power, followed by the two extended double ended colormaps (BOD and ECW). The rest of the colormaps (CW, VI, GP, RA) are indistinguishable at some sizes but not at others.

From Fig. 2 we can expect that discriminative power for the chromatically varying colormaps will increase with feature size much more rapidly than for the luminance varying colormap. To better understand the effects of feature size on the contrast thresholds a subset of the colormaps are replotted in Fig. 9. This shows data obtained with the rainbow (RA), gray (GP), blue-red (BY) and green-red (GR) colormaps with curves for the different sizes on each plot. The Rainbow colormap exhibits a large effect of feature size in the section to the left of center. This section represents the cyan to green range, where there is very little luminance variation, but large chromatic variation. The two areas at the ends of the measured section of the rainbow colormap have considerable luminance variation and the contrast sensitivity varies much less as a function of feature size. The blue-red (BY) and green-red (GR) colormaps also shows greater discriminative power for the 45 pixel feature sizes in comparison with the 10 and 15 pixel feature sizes, whereas there is very little variation in the gray scale (GP) in the sensitivity with respect to size.

To statistically test the hypothesis that the relative discriminative power of the non-luminance components of colormaps increase with feature size, we ran a two way ANOVA (colormap, feature size) on the results from the gray, green-red and blue-red color maps (GP,GR,BY). Both main effects and the interaction were highly significant. For colormap ($F(2,183) = 1213$, $p < 0.001$); for feature size ($F(2,183) = 23.95$, $p < 0.001$); for the interaction $F(4,183) = 31.75$, $p < 0.001$). Fig. 10 illustrates the interaction, showing the ratio of the average contrast sensitivities for the GR and BY colormaps with the gray (GP) colormap for each of the sizes. The results show that participants became increasingly sensitive to chromaticity differences (relative to luminance differences) as feature sizes increased.

6 Modeling the Results with CIELAB

As a step towards an engineering model of the kind developed and expanded upon in [51] and [52] for discrete colors we were interested in determining the extent to which a modified version of a UCS could account for our results. We began by investigating modified and unmodified versions of both CIELAB's CIEDE1976 and CIEDE2000. However, since CIELAB's euclidean metric (CIEDE1976) produced the best results in preliminary work and is considerably simpler, we conducted most of the analysis with standard CIELAB. The CIEDE2000 results were very similar but provided slightly lower correlation with the experimental data. We only present CIEDE1976 results here.

To fit the results, a set of intervals spanning each of the 30 test points was defined on the test colormaps. The colormap R,G,B values were converted to CIE X,Y,Z and then to (modified) CIELAB values assuming the sRGB standard. The CIELAB reference white, L_n was defined by $R = G = B = 1$. Color difference values $\log(\Delta E)$ were computed for these intervals for all nine sequences yielding a total of 270 values. These correspond to the 270 average log contrast sensitivity measurements obtained for each of the 10, 15, and 45 pixel patterns. To determine if changing weights on the CIELAB a^* and b^* terms more accurately account for the data, we computed the entire set of $\log(\Delta E)$ values for a matrix of weights on a^* and b^* values declining in steps of 0.05 using Equation (4). We also tested the fit using weights of 0.125, 0.075, and 0.025. The quality of the model fit was evaluated by calculating the r^2 correlations between the 270 measured data points (30 points on 9 colormaps) and the corresponding intervals in the modified CIELAB model.

$$\Delta E = \sqrt{(\Delta L^*)^2 + (w_a \Delta a^*)^2 + (w_b \Delta b^*)^2}. \quad (4)$$

TABLE 2
The r^2 Values for Regressions of Unweighted CIELAB (Second Column), Luminance Only (Third Column) and Weighted CIELAB (Fourth Column) against the Observed Resolving Power Results Along with the Best Fit Weights and Coefficients

Feature Size	Fit r^2: w_a,w_b = 1	Fit r^2 w_a,w_b = 0	Best fit r^2	Weight a	Weight b	Coeff α	Coeff β
10 pixels	0.386	0.819	0.961	0.075	0.075	0.867	0.517
15 pixels	0.438	0.861	0.975	0.075	0.075	0.874	0.572
45 pixels	0.521	0.831	0.970	0.125	0.125	0.799	0.730
Overall	0.429	0.815	0.940	0.100	0.100	0.879	0.531

The regression equation compared the average results obtained (recall that these are already on a log scale) with the log of the $\Delta E/\Delta s$ ratio for all 270 points.

$$log_{10}(c) = \alpha log_{10}(\Delta E/\Delta s) + \beta. \quad (5)$$

In Table 2, we show the r^2 values obtained with unmodified CIELAB ($w_a, w_b = 1$) and for CIELAB L* only ($w_a, w_b = 0$) as well as the best fits and their corresponding weights.

The regression parameters can be used together with the a^*, b^* weights to construct regression model-based curves corresponding to the measured contrast sensitivity curves. The result for the 15 and 45 pixel data is shown in Fig. 11. As can be seen, except for a few excursions, the fits are excellent when reduced weights on a^* and b^* are used. The results for the 10 pixel data are very similar to the 15 pixel results.

It is probably never the case that scientific data is made up of a single spatial frequency and for this reason, we also did regression fits using modified a^*, b^* weights to the combined data set including all three spatial frequencies. The result was an r^2 value of 0.94, with a best fit occurring with a^*, b^* weights = 0.1. The fit is given by the equation:

$$log_{10}(c) = 0.879 log_{10}(\Delta E/\Delta s) + 0.531, \quad (6)$$

where c is the measured contrast sensitivity, ΔE is the modified CIELAB value, and Δs is the interval. This can be rearranged to become

$$c = 3.4(\Delta E/\Delta s)^{0.879}. \quad (7)$$

To test whether the best CIELAB fits obtained with weighted a^* and b^* were better than the fits obtained with $w_a, w_b = 1$ we used an F test using the ratio of the regression residuals $(1 - r^2)$ [20] for each of the values in Table 2. The degrees of freedom are the number of data points minus the model degrees of freedom (270-2) for both the numerator and the denominator. The results from applying this test for all sizes and for the combined data, summarized in Table 3, show that the weighted model fits were significantly better than the results obtained with standard weights $w_a, w_b = 1$ with $p < 0.001$. In addition the weighted model fits were significantly better compared to L^*-only, at the $p < 0.001$ level, for all sizes and for the combined data.

We can also compare the path lengths of colormaps against the measured discriminative power shown in Fig. 8. Fig. 12 shows a comparison of the path length in both unmodified CIELAB and weighted CIELAB (w_a,w_b = 0.1) for only the measured section of the colormaps (between 0.1 and 0.9). As can be seen, the correlation is less than 0.8 for unmodified CIELAB and greater than 0.99 for modified CIELAB.

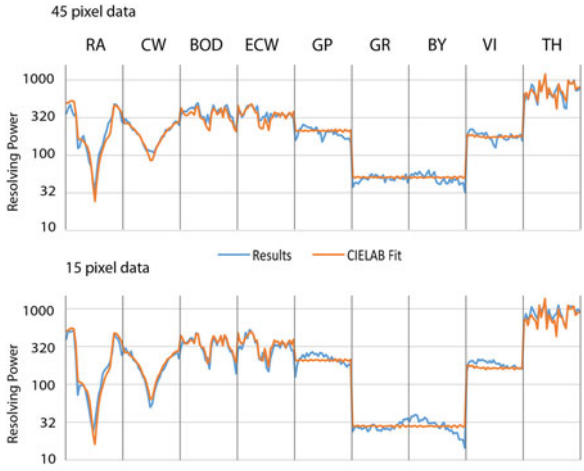

Fig. 11. The model fit to all nine colormaps for the (top) 45 pixel data and (bottom) 15 pixel data.

7 Discussion

Our original hypothesis was that using a uniform color space to predict the feature detection functions of colormaps will be inaccurate because these models give too much weight to chromatic channel information when smaller features are considered. Our new feature resolution method applied in a Mechanical Turk study yielded 270 average feature detection measurements for three different spatial frequencies. As hypothesized, unmodified CIELAB provided a poor model for the results. A much better model was obtained by greatly reducing the weights on the model terms corresponding to the green-red and yellow-blue color channels. We also tested against the hypothesis that L^* by itself could account for the data equally well and found that it could not, although this did better than unmodified CIELAB.

Overall, our results are in rough agreement from what would be expected from Fig. 2. At 3 cycles/deg. Mullen's results [32] show a ratio of approximately 5:1 between the contrast sensitivity of the color channels and the luminance channel. We found the contribution of the color channels to be even smaller than this. In addition, Fig. 2 shows the difference between color channel sensitivity and luminance sensitivity declining as spatial frequency decreases, something we also found. It will be interesting in future work to determine whether this trend continues for still larger patterns. However, a methodology other than the one we use here will be required.

The stimulus patterns we have developed provide an easy-to-use method for directly measuring the feature detection functions of colormaps [59]. But given the excellent results obtained with the modified CIELAB model, the model expressed in Equation (7) can be used as an

TABLE 3
F Tests Comparing Best Fits (From Table 2) with Fit Using Unmodified CIELAB (w_a,w_b = 1) and with L* Only (w_a,w_b = 0)

Feature Size	Test against unmodified CIELAB: w_a,w_b = 1	Test against Luminance only: w_a,w_b = 0
10 pixel	F(268, 268) = 15.74, $p < 0.001$	F(268, 268) = 4.67, $p < 0.001$
15 pixel	F(268, 268) = 22.48, $p < 0.001$	F(268, 268) = 5.56, $p < 0.001$
45 pixel	F(268, 268) = 15.97, $p < 0.001$	F(268, 268) = 5.63, $p < 0.001$
Overall	F(268, 268) = 9.51, $p < 0.001$	F(268, 268) = 3.08, $p < 0.001$

alternative. Fig. 13 shows a color sequence design tool we have constructed for this purpose. The left hand panel shows a slice through CIELAB space at a particular luminance value. While the best model fits were obtained by using different weights for the different pattern sizes, good correlations can be obtained using a single pair of weights. Choosing 10 percent weights on the a^* and b^* terms, for instance, would be a reasonable option for smaller feature sizes ($> 1 cycle/deg$). If somewhat larger features are of interest then weights of 15 or 20 percent may also be used.

The simple model expressed in Equation (7) should be only regarded as the first step towards a more complete engineering model. It has a number of shortcomings. First, while covering a broad range of color space, the set of colormaps on which it was based do not provide a systematic or uniform sampling of color space, and may not cover all regions of color space most likely to be used in creating colormaps. Second, it is not based on a systematic or uniform sampling of spatial frequency. Third, a more general model should include data on low spatial frequency patterns. Nevertheless, because the r^2 values varied smoothly and gradually over wide range of weights (which is why we were able to produce a respectable fit to all of the data with a single pair of weights) we believe that the model proposed here has value as a rough approximation until a more complete model becomes available.

The results reinforce the importance of luminance variation in the representation of features in data as already noted by prior researchers [42], [58]. Because of the minor contribution of color differences to feature detection a simple rule of thumb: "use lightness variation for pattern perception" still holds. Also, they explain why the Samsel BOD and ECW colormaps provide great discriminative power; it is because they substantially increase the pathlength, especially with respect to luminance, over colormaps which vary monotonically in luminance such as Viridis. The thermal imaging colormap provides an extreme example of this. It has more than four times the discriminative power of Viridis.

The results presented here only apply to features >1 cycle/degree of visual angle. We do not know the extent to which important features in scientific data fall in this size range, but it may well be the majority because far more information can be conveyed with high spatial frequency channels than low spatial frequency channels. The amount of information carried on a channel varies with the frequency [48]. For two dimensional patterns this becomes the square of the frequency. Based on the human spatial modulation sensitivity function provided in Watson [60], hundreds of times more perceivable information can be carried at spatial frequencies above 1 cycle/degree than can be carried at lower spatial frequencies.

Resolution of constituent features provides a necessary condition for pattern perception, but it is far from being the entire story. The perception of features that are well above detection threshold almost certainly depends on a number of additional perceptual mechanisms relating to contour perception and shape perception. But these too are likely to depend mostly on luminance variations. Certainly this is true for faces in the work of [14], [19], [42].

We wish to be clear that we are not advocating the use of double ended colormaps for most cases. In general it is better to reserve the use of double ended colormaps for cases where values vary above and below some baseline, as is commonly done in the case of temperature anomalies. Nevertheless, it is the case that the extended double-ended colormaps do offer greater ability to resolve features, and where this is a critical requirement they can be valuable for this purpose.

There is also the issue of consistency in luminance variation. There is a cost to changing the direction of luminance variation within a colormap as Fig. 14 illustrates. The ring patterns in that image are consistently visible for Viridis and green-red. With the ECW colormap, they are light on the left, but dark on the right and this is confusing. The thermal imaging colormap has multiple zigzags with respect to luminance and this makes it extremely confusing (see also

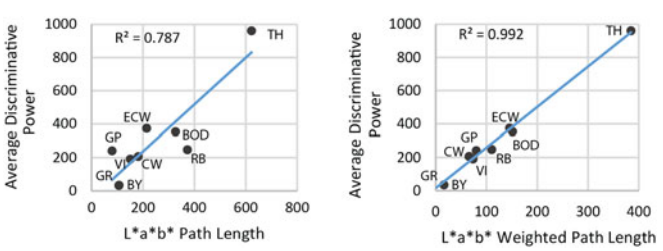

Fig. 12. A comparison of the average discriminative power as a function of the colormap path length in (left) CIELAB and (right) the weighted CIELAB. Note the improved fit using the weighted model.

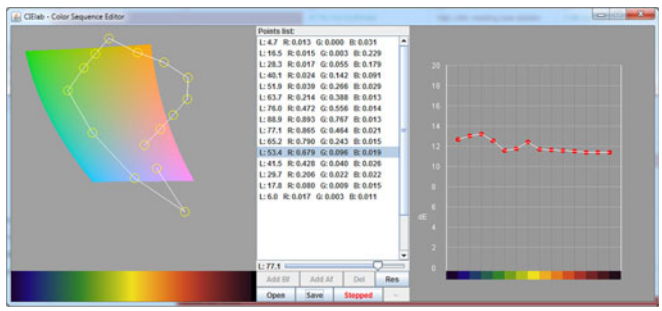

Fig. 13. A color sequence design tool. The left hand panel shows a slice through CIELAB at the luminance level of the selected point. The plot on the right shows the feature detection function based on the modified CIELAB model. A simple double ended, uniform rainbow colormap has been constructed.

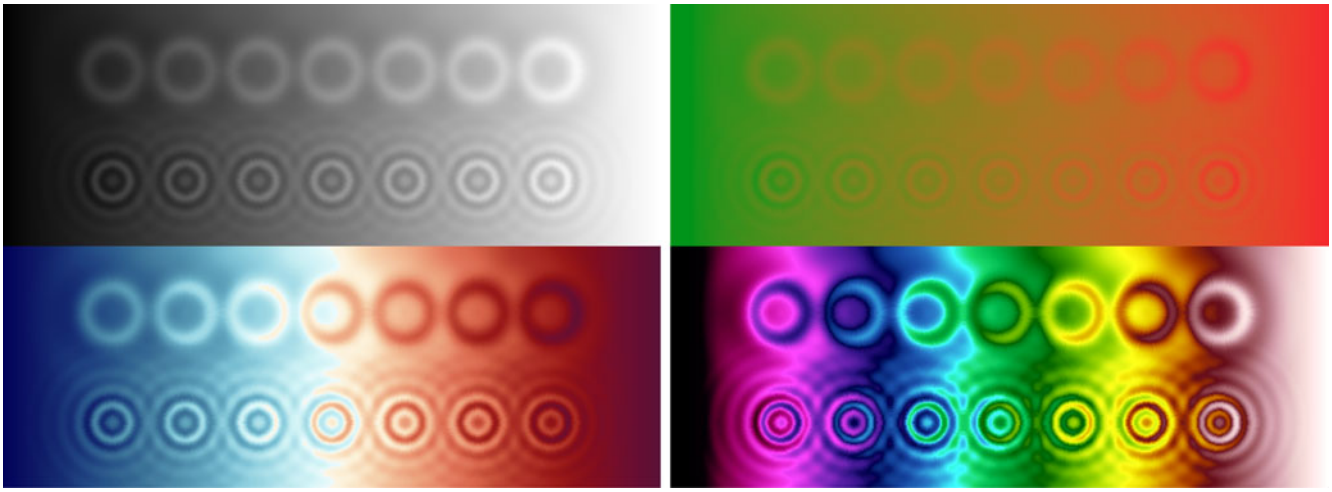

Fig. 14. In this figure, artificial data has a background ramp increasing from the left. Superimposed on the ramp are ripple patterns. Top: grayscale (left), green/red (right); bottom: extended cool/warm (left), thermal (right).

Fig. 1). The only case where we can imagine that it may be useful is where extreme feature resolution is critical. This is presumably why it is sometimes used.

The results also confirm prior work that has shown chroma scales to have fewer resolvable steps compared to luminance scales [4], [5], although they are perceptually ordered. If a chroma scale is chosen, it will likely benefit from some small amount of monotonic luminance variation in addition to the progressive increase in chroma.

Our results are qualitatively similar to those of [43] and [17] even though our method is very different. They similarly found a very sharp peak in contrast threshold (the inverse of contrast sensitivity) in the center of the rainbow colormap. They noted although chroma scales appear to be good candidates for creating colormaps for encoding data magnitude they had lower contrast sensitivities compared to gray scales.

7.1 Discussion on Crowdsourcing

Crowdsourcing user evaluation has become increasingly common over the past decade, including within the visualization community. While there are obviously trade-offs between the ecological validity available with a wide demographic cross-section versus the level of experimental control, the community is rapidly lining up on the side of easy participant recruitment and quick turn-around time.

Mturk does have the potential for contamination due to color vision deficiencies. Approaches to minimize CVD contamination range from simply asking people to self-select for a given study, (e.g., Szafir's recent Best Paper at InfoVis 2017 [52]) to including a CVD test such as Ishihara plates (e.g., Liu and Heer [23]) although as the authors acknowledge, online presentation of Ishihara plates has potential pitfalls (due to, e.g., unknown monitor calibration or allowing participants infinite time to respond).

Research on crowdsourced participant pools [13], [45] has shown that while participants are very consistent in their demographic responses across many studies (e.g., 98.9 percent gender consistency), they are indeed more likely to lie when a lucrative reward is offered but restricted to a certain demographic. Our approach to minimize potential contamination due to CVD [56] exploits these tendencies by periodically launching a study to sweep self-identified participants with CVD into an exclusion pool. The study specifically requests participants with CVD and presents participants with a valid test for CVD. Anyone taking the study, either colorblind or potentially lying to garner the fee, are put into the exclusion group. The result is an Mturk participant pool with, not the expected ≈4.5 percent occurrence of CVD in the general population, but something much less, albeit unspecified. Additionally, during an actual study, participants are asked for their CVD status and removed if they have CVD.

We can again validate this approach by comparing the male/female response within a subset of this current study. The three colormaps of particular interest for CVD are the grayscale (GP), the green-red (GR) and the blue-red (YB). The data for these three colormaps was gathered via a within-subject study. Given the very low occurrence of CVD in women, we used the TurkPrime [22] *gender consistency score* to separate the participants into male and female, requiring a gender consistency score of 100 percent. Note that not all participants have a calculated gender consistency score as it is not assigned for participants with fewer than 100 studies launched on TurkPrime. Given the results of Fig. 9, we combined the 10 pixel and 15 pixel data to increase statistics and simply summed the raw vertical pixel response for all 30 data points for each participant. An independent two-sample t-test was conducted to compare raw pixel response for male (N = 13) and female (N = 21). We found no significant difference for any of the three colormaps as summarized in Table 4.

TABLE 4
Summary of Independent T-Test for Male and Female Response Across the Three Colormaps of Interest for CVD

Colormap:	Grayscale (GP)	Green-Red (GR)	Blue-Red (BY)
Mean (M)	6717	10900	11281
Std Dev (M)	1737	10768	11491
Mean (F)	6308	1875	1908
Std Dev (F)	1078	1118	1263
p (two-tail)	0.46	0.82	0.73

Anecdotally, over years of crowdsourcing color studies, we have found that CVD participants often appear as outliers in the data. Hence, a common-sense approach of asking people to self-identify in combination with effective data-scrubbing is probably sufficient to mitigate the risk of CVD contamination for typical crowdsourced experiments involving color.

Likewise, since this study was carried out using Mechanical Turk, certain caveats apply relating to the use of that platform. The laptop or desktop screens viewed by study participants were almost certainly not calibrated and the resolutions were unknown. Because of this we can only give an estimate of the actual spatial frequencies of the test patterns. Yet, despite these limitations, we were able to obtain remarkably clean data. The great advantage of a Mechanical Turk study in applied research such as this is the ecological validity. Compare our study with hundreds of participants with the psychophysical studies on which spatial color theory is based; the latter had only only one or two participants [32], [36]. The goal is to produce colormaps that are effective under a range of viewing conditions and across many scientists. For this reason, the variety of both monitors and study participants is a major asset.

8 Conclusion

The method we have developed provides a simple and quick way of evaluating the uniformity of colormaps. It produces remarkably consistent results, even with a study environment that lacks the normal laboratory controls for user studies.

The work with CIELAB modifications provides a link between spatial vision, color theory and practical problems of colormap design. The results and theory both suggest that colormap uniformity is not a simple concept, since the relative weights of chromatic variation and luminance variation change as a function of the spatial frequency of features. Nevertheless, a simple modification to CIELAB can produce a far better model for the detection of patterns in colormapped data where those patterns are composed of features with spatial frequencies of one cycle per degree and higher.

Acknowledgments

This material is based upon work supported by Dr. Lucy Nowell of the U.S. Department of Energy Office of Science, Advanced Scientific Computing Research under Award Numbers DE-AS52-06NA25396, DE-SC-0012438, and DE-SC-0012516. Funding was also provided under NOAA grant NA15NOS4000200 to the UNH Center for Coastal and Ocean Mapping. The authors would like to thank Dr. James Ahrens. This work was released under LA-UR-18-21476.

References

[1] J. Ahrens, B. Geveci, C. Law, C. Hansen, and C. Johnson, "ParaView: An end-user tool for large-data visualization," *Visualization Handbook*, pp. 717–731, 2005.

[2] Amazon Mechanical Turk Website. [Online]. Available: www.mturk.com/mturk/welcome

[3] S. Antis and P. Cavanagh, "A minimum motion technique for judging equiluminance", in *Colour Vis.: Physiology and Psychophysics*. New York NY, USA: Academic Press, pp. 155–166, 1983.

[4] L. D. Bergman, B. E. Rogowitz, and L. A. Treinish, "A rule-based tool for assisting colormap selection," in *Proc. 6th Conf. Vis.*, 1995, Art. no. 118.

[5] J. Bertin, *Semiology of Graphics: Diagrams, Networks, Maps*. Madison, WI, USA: University of Wisconsin Press, 1983.

[6] C. Blakemore and F. W. Campbell, "On the existence of neurones in the human visual system selectively sensitive to the orientation and size of retinal images," *J. Physiol.*, vol. 203, no. 1, pp. 237–260, 1969.

[7] D. Borland and A. Huber, "Collaboration-specific color-map design," *IEEE Comput. Graph. Appl.*, vol. 31 no. 4, pp. 7–11, Jul./Aug. 2011, doi: 10.1109/MCG.2011.55.

[8] D. Borland and R. M. Taylor II, "Rainbow color map (still) considered harmful," *IEEE Comput. Graph. Appl.*, vol. 27 no. 2, pp. 14–17, Mar./Apr. 2007, doi: 10.1109/MCG.2007.46.

[9] C. A. Brewer, "Color use guidelines for mapping," in *Visualization in Modern Cartography*, Oxford, U.K.: Pergammon, pp. 123–148, 1994.

[10] A. D. Broadbent, "Calculation from the original experimental data of the CIE 1931 RGB standard observer spectral chromaticity coordinates and color matching functions," Québec, Canada: Département de génie chimique, Université de Sherbrooke, 2008.

[11] R. Bujack, T. L. Turton, F. Samsel, C. Ware, D. H. Rogers, and J. Ahrens, "The good, the bad, and the ugly: A theoretical framework for the assessment of continuous colormaps," *IEEE Trans. Vis. Comput. Graph.*, vol. 24, no. 1, pp. 923–933, Jan. 2018, doi: 10.1109/TVCG.2017.2743978.

[12] F. Campbell and J. Robson, "Application of Fourier analysis to the visibility of gratings," *J. Physiol.*, vol. 197, no. 3, 1968, Art. no. 551566.

[13] J. J. Chandler and G. Paolacci, "Lie for a dime: When most prescreening responses are honest but most study participants are impostors," *Social Psychological Personality Sci.*, vol. 8, no. 5, pp. 500–508, 2017, doi: 10. 1177/1948550617698203.

[14] R. L. Gregory, "Vision with isoluminant colour contrast: 1. a projection technique and observations," *Perception*, vol. 6, no. 1, pp. 113–119, 1977, "PMID: 840617, doi: 10.1068/p060113.

[15] R. Huertas, M. Melgosa, and C. Oleari, "Performance of a color-difference formula based on OSA-UCS space using small-medium color differences," *J. Opt. Soc. Amer. A*, vol. 23, no. 9, pp. 2077–2084, 2006.

[16] L. M. Hurvich, *Color Vis*. Sunderland, MA, USA: Sinauer Associates Inc., 1981.

[17] A. D. Kalvin, B. E. Rogowitz, A. Pelah, and A. Cohen, "Building perceptual color maps for visualizing interval data," in *Proc. Human Vis. Electron. Imag. V*, 2000, vol. 3959, pp. 323–336.

[18] N. Kanwisher, J. McDermott, and M. M. Chun, "The fusiform face area: A module in human extrastriate cortex specialized for face perception," *J. Neurosc.*, vol. 17, no. 11, pp. 4302–4311, 1997.

[19] G. Kindlmann, E. Reinhard, and S. Creem, "Face-based luminance matching for perceptual colormap generation," in *Proc. Conf. Vis.*, 2002, pp. 299–306.

[20] M. H. Kutner, C. Nachtsheim, and J. Neter, *Applied Linear Regression Models*. New York, NY, USA: McGraw-Hill/Irwin, 2004.

[21] H. Levkowitz, "Perceptual steps along color scales," *Int. J. Imag. Syst. Technol.*, vol. 7, no. 2, pp. 97–101, 1996.

[22] L. Litman, J. Robinson, and T. Abberbock, "TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences," *Behavior Res. Methods*, vol. 49. pp. 1–10, 2016, doi: 10.3758/s13428-016-0727-z.

[23] Y. Liu and J. Heer, "Somewhere over the rainbow: An empirical assessment of quantitative colormaps," in *Proc. Conf. Human Factors Comput. Syst.*, 2018, pp. 598:1–598:12, doi: 10.1145/3173574.3174172.

[24] M. R. Luo, G. Cui, and B. Rigg, "The development of the CIE 2000 colour-difference formula: CIEDE2000," *Color Res. Appl.*, vol. 26, no. 5, pp. 340–350, 2001.

[25] M. R. Luo and C. Li, "CIECAM02 and its recent developments," in *Advanced Color Image Processing and Analysis*. Berlin, Germany: Springer, 2013, pp. 19–58.

[26] D. L. MacAdam, "Visual sensitivities to color differences in daylight∗," *J. Opt. Soc. Amer.*, vol. 32, no. 5, pp. 247–274, May 1942, doi: 10.1364/JOSA.32.000247.

[27] M. Mahy, L. Eycken, and A. Oosterlinck, "Evaluation of uniform color spaces developed after the adoption of CIELAB and CIELUV," *Color Res. Appl.*, vol. 19, no. 2, pp. 105–121, 1994.

[28] G. McCarthy, A. Puce, J. C. Gore, and T. Allison, "Face-specific processing in the human fusiform gyrus," *J. Cognitive Neuroscience*, vol. 9, no. 5, pp. 605–610, Oct. 1997, doi: 10.1162/jocn.1997.9.5.605.

[29] S. Mittelstädt, D. Jäckle, F. Stoffel, and D. A. Keim, "ColorCAT: Guided design of colormaps for combined analysis tasks," in *Proc. Eurographics Conf. Vis.*, 2015, pp. 115–119, doi: 10.2312/eurovisshort.20151135.
[30] K. Moreland, "Diverging color maps for scientific visualization," in *Proc. Int. Symp. Vis. Comput.*, 2009, pp. 92–103.
[31] N. Moroney, M. D. Fairchild, R. W. Hunt, C. Li, M. R. Luo, and T. Newman, "The CIECAM02 color appearance model," in *Proc. Color Imag. Conf.*, 2002, pp. 23–27.
[32] K. T. Mullen, "The contrast sensitivity of human colour vision to red-green and blue-yellow chromatic gratings," *J. Physiol.*, vol. 359, pp. 381–400, 1985.
[33] I. C. on Illumination, "Colorimetry,"Commission internationale de l'Eclairage, CIE Central Bureau, *CIE Tech. Rep.*, 2004, https://books.google.com/books?id=P1NkAAAACAAJ
[34] L. Padilla, P. S. Quinan, M. Meyer, and S. H. Creem-Regehr, "Evaluating the impact of binning 2d scalar fields," *IEEE Trans. Vis. Comput. Graph.*, vol. 23, no. 1, pp. 431–440, Jan. 2017, doi: 10.1109/TVCG.2016.2599106.
[35] S. M. Pizer, "Intensity mappings to linearize display devices," *Comput. Graph. Image Process.*, vol. 17, no. 3, pp. 262–268, 1981.
[36] A. B. Poirson and B. A. Wandell, "Pattern-color separable pathways predict sensitivity to simple colored patterns," *Vis. Res.*, vol. 36, no. 4, pp. 515–526, 1996, doi: 10.1016/0042–6989(96)89251-0.
[37] Qualtrics Website. [Online]. Available: www.qualtrics.com
[38] P. L. Rheingans, "Task-based color scale design," in *Proc. 28th AIPR Workshop: 3D Vis. Data Exploration Decision Making*, 2000, pp. 35–43.
[39] A. R. Robertson, "The CIE 1976 color-difference formulae," *Color Res. Appl.*, vol. 2, no. 1, pp. 7–11, 1977, doi: 10.1002/j.1520–6378.1977.tb00104.x.
[40] P. K. Robertson and J. F. O'Callaghan, "The generation of color sequences for univariate and bivariate mapping," *IEEE Comput. Graph. Appl.*, vol. 6, no. 2, pp. 24–32, Feb. 1986.
[41] P. K. Robertson and J. F. O'Callaghan, "The application of perceptual color spaces to the display of remotely sensed imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 26, no. 1, pp. 49–59, Jan. 1988, doi: 10.1109/36.2999.
[42] B. E. Rogowitz and A. D. Kalvin, "The "Which Blair Project": A quick visual method for evaluating perceptual color maps," in *Proc. Vis.*, 2001, pp. 183–556.
[43] B. E. Rogowitz, A. D. Kalvin, A. Pelah, and A. Cohen, "Which trajectories through which perceptually uniform color spaces produce appropriate colors scales for interval data?" in *Proc. 7th Color Imag. Conf.*, 1999, pp. 321–326.
[44] B. E. Rogowitz and L. A. Treinish, "Data visualization: The end of the rainbow," *IEEE Spectrum*, vol. 35, no. 12, pp. 52–59, Dec. 1998.
[45] The TurkPrime Team, "Are MTurk workers who they say they are?," *Effective Mech. Turk: The TurkPrime Blog*, 1 Dec. 2017, http://blog.turkprime.com/2017/12/are-mturk-workers-who-they-say-they-are.html
[46] F. Samsel, M. Petersen, T. Geld, G. Abram, J. Wendelberger, and J. Ahrens, "Colormaps that improve perception of high-resolution ocean data," in *Proc. 33rd Annu. ACM Conf. Extended Abstracts Human Factors Comput. Syst.*, 2015, pp. 703–710, doi: 10.1145/2702613.2702975.
[47] F. Samsel, T. L Turton, R. Bujack, D. H. Rogers, J. Ahrens, G. D. Abram, and C. Ware, Data Science at Scale Sci-Vis color website, https://sciviscolor.org
[48] C. E. Shannon, A. Wyner, and N. J. Sloane, "*Claude E. Shannon: Collected Papers*," Hoboken, NJ, USA: Wiley, 1993.
[49] S. S. Stevens, *Psychophysics*. Piscataway, NJ, USA: Transaction Publishers, 1975.
[50] M. Stone, *A Field Guide to Digital Color*. Boca Raton, FL, USA: CRC Press, 2016.
[51] M. Stone, D. A. Szafir, and V. Setlur, "An engineering model for color difference as a function of size," in *Proc. 22nd Color Imag. Conf. Final Program Proc.*, 2014, pp. 253–258.
[52] D. A. Szafir, "Modeling color difference for visualization design," *IEEE Trans. Vis. Comput. Graph.*, vol. 24, no. 1, pp. 392–401, Jan. 2018, doi: 10.1109/TVCG.2017.2744359.
[53] J. Tajima, "Uniform color scale applications to computer graphics," *Comput. Vis., Graph. Image Process.*, vol. 21, no. 3, pp. 305–325, 1983.
[54] C. Tominski, G. Fuchs, and H. Schumann, "Task-driven color coding," in *Proc. 12th Int. Conf. Inform. Vis.*, 2008, pp. 373–380.
[55] B. E. Trumbo, "A theory for coloring bivariate statistical maps," *Amer. Statistician*, vol. 35, no. 4, pp. 220–226, 1981.
[56] T. L. Turton, C. Ware, F. Samsel, and D. H. Rogers, "A crowdsourced approach to colormap assessment," in *EuroVis Workshop on Reproducibility, Verification, and Validation in Visualization (EuroRV3)*, L. Kai, S. Noeska, and C. Douglas Eds., The Eurographics Association, 2017, doi: 10.2312/eurorv3.20171106.
[57] S. van der Walt and N. Smith, "Matplotlib documentation update." [Online]. Available: bids.github.io/colormap/
[58] C. Ware, "Color sequences for univariate maps: Theory, experiments and principles," *IEEE Comput. Graph. Appl.*, vol. 8, no. 5, pp. 41–49, Sep. 1988.
[59] C. Ware, T. L. Turton, F. Samsel, R. Bujack, and D. H. Rogers, "Evaluating the perceptual uniformity of color sequences for feature discrimination," in *Proc. EuroVis. Workshop Reproducibility Verification Validation Vis.*, 2017, pp. 1–5, doi: 10.2312/eurorv3.20171107.
[60] A. B. Watson, "Visual detection of spatial contrast patterns: Evaluation of five simple models," *Opt. Exp.*, vol. 6, no. 1, pp. 12–33, 2000, doi: 10.1364/OE.6.000012.
[61] M. Wijffelaars, R. Vliegen, J. J. Van Wijk, and E.-J. Van Der Linden, "Generating color palettes using intuitive parameters," *Comput. Graph. Forum*, vol. 27, no. 3, pp. 743–750, 2008, doi: 10.1111/j.1467-8659.2008.01203.x.
[62] L. Zhou and C. Hansen, "A survey of colormaps in visualization," *IEEE Trans. Vis. Comput. Graph.*, vol. 22, no. 8, pp. 2051–2069, Aug. 2016.

Colin Ware He received the PhD degree in psychology from the University of Toronto, in1980, and the MMath degree in computer science from the University of Waterloo, in 1985. He is the director of the Visualization Research Lab in the Center for Coastal and Ocean Mapping, University of New Hampshire. With a cross appointment between the Departments of Ocean Engineering and Computer Science, he has a special interest in the application of perceptual theory to visualization design and human-computer interaction. He has published more than 150 articles in scientific and technical journals and leading conference proceedings. He has also written two books: *Visual Thinking for Design* is an account of the psychology of how we think using graphic displays as tools. *Information Visualization: Perception for Design*, now in its 3rd edition, is a comprehensive survey of what human perception tells us about how to display information.

Terece L. Turton received the BS degree with University Honors in physics from Carnegie Mellon University, in 1988, the MS and PhD degrees in physics from the University of Michigan, in 1988 and 1993, respectively. She joined the Data Science at Scale Team at Los Alamos National Laboratory, in 2018 after three years with the University of Texas at Austin. She worked on various high energy physics experiments. Her current research interests include perceptual user evaluation and workflow analysis in scientific visualization.

Roxana Bujack received the Diplom (MS) in mathematics, in 2010, the BSc degree in computer science, in 2011, and the PhD degree summa cum laude in computer science from Leipzig University, in 2014. She joined the Data Science at Scale Team at Los Alamos National Laboratory, in 2016. Her research interests include visualization, pattern detection, vector fields, moment invariants, high performance computing, massive data analysis, Lagrangian flow representations, and Clifford analysis.

Francesca Samsel She received the BFA degree from the California College of Art, and the MFA degree from the University of Washington. Multidisciplinary collaboration between the arts, sciences and visualization form the core of her research, specifically, identifying artistic principles and expertise with potential to assist scientists in their scientific inquiries. She is a research associate with the Center for Agile Technology, University of Texas at Austin. She works in collaboration with computational teams at Los Alamos National Laboratory, the Texas Advanced Computing Center, and the University of Minnesota, Interactive Visualization Laboratory. She is a member of the Editorial Board of IEEE Computer Graphics and Application, as well as on the program committee for IEEE Vis, Visual Arts Program

Piyush Shrivastava received the BEng degree in computer science from Rajiv Gandhi Technical University. He is working toward the graduate degree in the Computer Science Department, University of New Hampshire. He has five years of experience as a software engineer.

David H. Rogers received the degree in computer science from the University of New Mexico, in 1996, the degree in architecture, Princeton, 1988, and the MFA degree in writing for children. He joined Los Alamos National Laboratory, in 2013, after a decade of leading the Scalable Analysis and Visualization Team at Sandia National Laboratory. Currently the team lead for the Data Science at Scale team with LANL, he now focuses on interactive web-based analysis tools that integrate design, scalable analytics, and principles of cognitive science to promote scientific discovery. Prior to working on large scale data analysis, he worked with DreamWorks Feature Animation.