Supporting Information: Assessing Environmental Oil Spill Based on Fluorescence Images of Water Samples and Deep Learning

Dongpeng Liu[† a], Ming Liu[†‡ a], Guangyu Sun[†], Zhiqian Zhou[†], Duolin Wang[†], Fei He[†], Jiaxin Li[†], Jiacheng Xie[†], Ryan Gettler[¶], Eric Brunson[¶], Jeffery Steevens[¶b], and Dong Xu[†b]

†*Department of Electrical Engineering and Computer Science, Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, Missouri 65211, USA.*

‡ *School of Mathematics and Statistics, Changchun University of Technology, Changchun, 130024, China.*

¶*U.S. Geological Survey Columbia Environmental Research Center, Columbia, Missouri 65201, USA.*

[a]Equal Contributions.

[b]Corresponding Authors: jsteevens@usgs.gov, xudong@missouri.edu

Disclaimer: Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government."

Feature Analysis

Color Representation

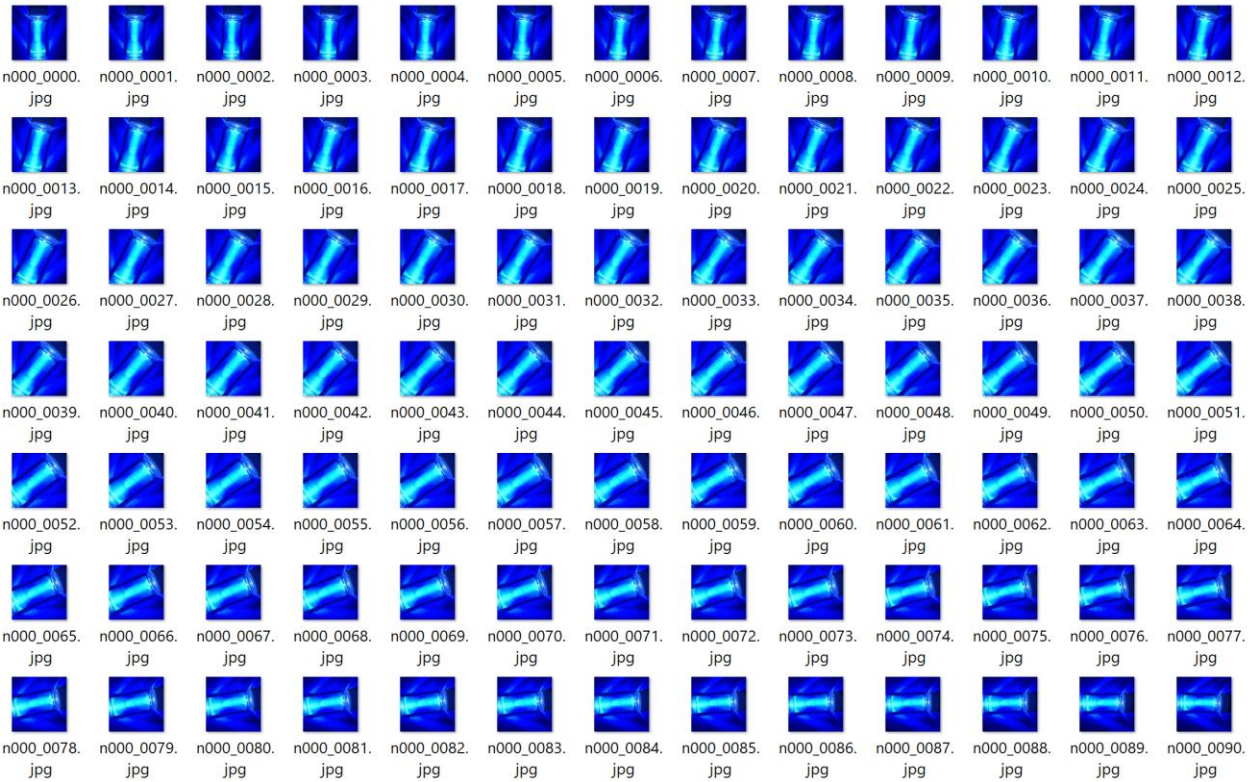

Figure S1: An example of the up-sampling protocol, where 360 pictures were generated from one original image.

Each image was up-sampled to 360 pictures, as shown in Fig. S1, for modeling training. For each picture, since color is the most intuitive and vital feature in our dataset, we studied color encoding in detail. An apparent visual feature of our image dataset is object free (low-frequency, containing only pure color), as shown in Figure 2. As human eyes cannot tell the difference between images, we performed statistical analyses by hierarchically examining our data:

1. Data feature in the whole dataset
    (a) Similarities measured by the distribution of histogram similarities.
    (b) Distribution of the maximum values in the red, green, blue (RGB) channels separately.
2. Data feature per image (random sampling some image for analysis)
    (a) Histogram and hue, saturation and value (HSV) analysis.
    (b) Frequency analysis; Gradient analysis: in x and y directions.

Histogram Similarity Analysis

We explored Pearson correlation coefficients of different features, including histogram similarity with and without spatial information, structural similarity (SSIM), light similarity, blue channel similarity, and gradient similarity.

Table S1: Pearson correlation coefficients of different features.

Feature	Pearson correlation coefficient
Histogram Similarity w/o spatial Information	-0.31
Histogram Similarity w/ spatial Information	-0.50
Structural Similarity	-0.27
Light Similarity	0.15
Blue Channel Similarity	0.10
Gradient Similarity	0.11

We chose $|labelx - labely|$ as the metric for label similarity and selected $\frac{|hist_x - hist_y|}{max(hist_x, hist_y)}$ as the metric for histogram similarity, $|lightx - lighty|$ as the metric for light similarity, $|bluex - bluey|$ as the

metric for blue channel similarity, and |*gradientx* − *gradienty*| as the metric for gradient similarity. Table S1 shows that histogram similarity is the most correlated feature to the label (the oil concentration).

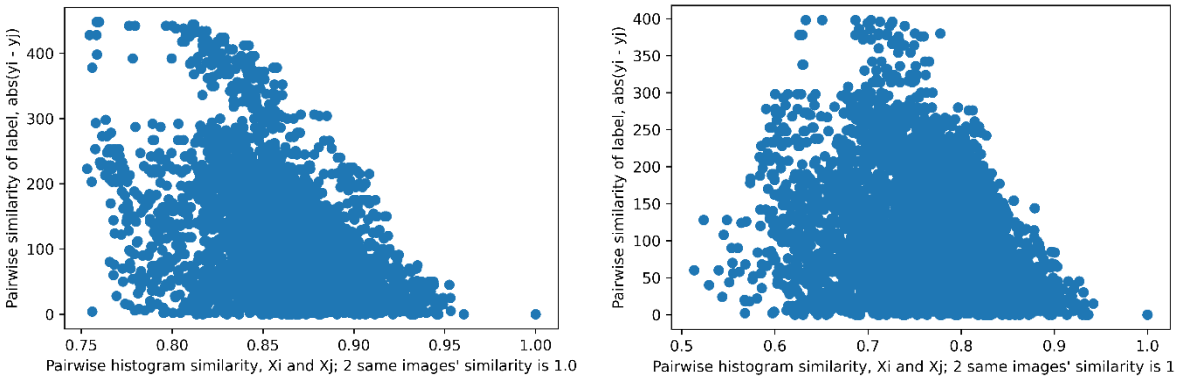

Figure S2: Similarity of the histogram with respect to the difference in the label. Histogram with (left) and without (right) the spatial information both show a specific correlation with the label difference.

Furthermore, two different histogram approaches were tested to get the relationship of histogram similarity with respect to the difference in labels. The first one is to compute the similarity of the whole image directly so that the similarity only carries color information. The other is to divide the image into 16 parts of the same size, then compute the average similarity of the 16 parts so that the final histogram contains the spatial information. Figure S2 shows that the split image carrying the spatial information has a higher correlation than the whole image with only the color information. Therefore, the spatial color feature is a better choice than the color only.

RGB Channel Analysis

To explore which color can play the most critical role in the prediction, we divided each image into three channels: Red (R), Green (G), and Blue (B). We compared each channel's maximum color intensity (the richest bucket), denoted argmax_arr, and the amount of nonzero-intensity (empty buckets) of its histogram, denoted count_nonzero_arr for each color. Figure S3 shows RGB color features and label correlation matrix. The blue feature shows the strongest correlation with the label among the three colors. For each image, we also recorded the most frequent color intensity in each channel and analyzed which color channel had the highest expectation of intensity. The kernel density estimation (KDE) plot is shown in Figure S4(a). It shows that the blue channel has the highest intensity. We also drew the KDE plot for nonzero intensity, as shown in Figure S4(b). The amount of nonzero intensity reflects the intensity variations of the channel. The blue channel also has the highest mean amount of nonzero intensity. As a result, we can assume that the blue channel is the most informative in our dataset for the prediction. We further randomly selected three images with labels and drew their histograms, as shown in Figure S4(c), supporting this observation. In addition, hue, saturation, and value (HSV) analysis shows that most images share similar ranges of hue (≈ 222), saturation (≈ 0.79), and brightness (≈ 0.81). Therefore, we can assume that all images are collected under a stable light source.

	argmax_arr_R	count_nonzero_arr_R	argmax_arr_G	count_nonzero_arr_G	argmax_arr_B	count_nonzero_arr_B	selected_y
argmax_arr_R	1.00	-0.18	0.62	-0.22	0.12	0.01	0.17
count_nonzero_arr_R	-0.18	1.00	-0.38	0.67	-0.57	0.51	-0.11
argmax_arr_G	0.62	-0.38	1.00	-0.38	0.37	-0.39	-0.30
count_nonzero_arr_G	-0.22	0.67	-0.38	1.00	-0.77	0.77	0.28
argmax_arr_B	0.12	-0.57	0.37	-0.77	1.00	-0.83	-0.34
count_nonzero_arr_B	0.01	0.51	-0.39	0.77	-0.83	1.00	0.57
selected_y	0.17	-0.11	-0.30	0.28	-0.34	0.57	1.00

Figure S3: RGB color features and label correlation matrix. argmax_arr represents the maximum color intensity (the richest bucket) and count_nonzero_arr represents the amount of nonzero-intensity (empty buckets) of its histogram. selected_y is the label. Deeper blue and red denote more significant correlations.

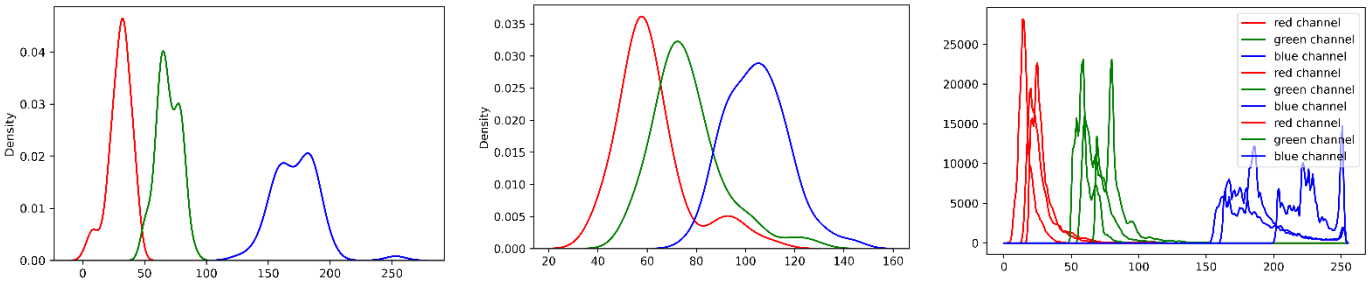

Figure S4. (a) E plot of the most frequent color intensity. (b) KDE plot of the range of RGB in the dataset. (c) RGB histograms on 3 with label 0, 26, 48 datasets. The horizontal axis shows the intensity value and the vertical axis indicates the density (distribution) of the intensity.

Frequency and Gradient Analysis

Figure S5 shows the analysis result of the fast Fourier transform (FFT) on a single image. It indicates that the whole image has not only non-segmentable objects but also no prominence gradient. Thus, the information of the image frequency spectrum is limited. We also analyzed images with different labels and their frequency features keep a high degree of similarity. It also shows that inside each feature map of the deep-learning model, the magnitude changing is flat, i.e., by random sampling, most feature maps have low variance, in other words, low-frequency variations, as the input images. In addition, feature maps near the input layer have lower variances than feature maps in deeper layers.

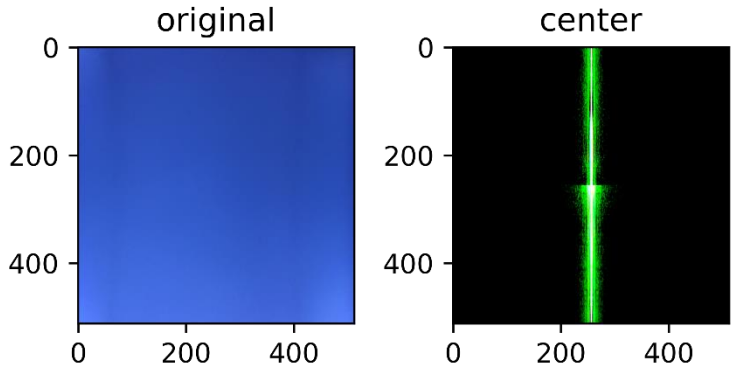

Figure S5: Image example and its fast Fourier transformation.