Learning to Detect Multiple Photographic Defects (Supplementary Material)

Ning Yu[1] Xiaohui Shen[2] Zhe Lin[2] Radomír Měch[2] Connelly Barnes[1]
[1]University of Virginia [2]Adobe Research
{ny4kt, connelly}@cs.virginia.edu {xshen, zlin, rmech}@adobe.com

1. Photographic Defect Severity Dataset

1.1. User Interface

Figure 1 left and bottom show an example of our Amazon Mechanical Turk (AMT) user interface. Figure 1 top right reports the corresponding severity ground truth for each of the seven defects averaged from five users.

1.2. Quality Control Schemes

In order to obtain the highest possible accuracy from AMT users, we incorporated two quality control mechanisms into the study.

Instruction. We showed users in AMT an instruction Web page with definitions (e.g. definition of "Exposure") and *none / mild / severe* examples for each defect (see Figure 2).

Qualification test. We additionally required users to pass a qualification test in AMT with 11 multiple-choice questions and 13 points in total (see Figure 3). The questions in the test are educational with obvious answers. Only users who passed the test with 11 points or higher can proceed to the real annotating tasks in AMT.

2. Simultaneous Detection of Multiple Defects

2.1. Defect-Specific Infogain Matrix Design

The infogain loss E is mathematically formulated as

$$E = -\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K} H_{l_n,k}\log(p_{n,k}), \quad (1)$$

where N is the number of image samples; K is the number of classes; l_n is the class ground truth of the n^{th} sample; $p_{n,k}$ is the probability of the n^{th} sample being classified as the k^{th} class, which is the output after the softmax layer satisfying $\sum_k^K p_{n,k} = 1$ and $p_{n,k} \geq 0$. Finally, $H_{l_n,k}$ is the infogain weight for the n^{th} sample with ground truth l_n to be classified to class k. The higher the weight, the greater the reward for that classification result.

Our design for the defect-specific infogain matrix $H = \{H_{l_n,k}\}$ is as follows. For a certain defect, suppose H is known and fixed, $\forall n \in \{1,...,N\}$ and $j \in \{1,...,K-1\}$, the optimal solution for $p^*_{n,j}$ satisfies

$$\begin{cases} \left.\frac{dE}{dp_{n,j}}\right|_{p^*_{n,j}} = -\frac{1}{N}\left(\frac{H_{l_n,j}}{p^*_{n,j}} - \frac{H_{l_n,K}}{1-\sum_{k=1}^{K-1} p^*_{n,k}}\right) = 0 \\ \sum_{k=1}^{K} p^*_{n,k} = 1, p^*_{n,j} \geq 0 \end{cases}. \quad (2)$$

One obvious solution of Equation (2) is

$$p^*_{n,j} = \frac{H_{l_n,j}}{\sum_{k=1}^{K} H_{l_n,k}}, \forall j \in \{1,...,K\}. \quad (3)$$

Since we have sufficient freedom to design H, we can set $\sum_{k=1}^{K} H_{l_n,k} = 1$ without loss of generality. If we make this assumption along with Equation (3), then we can derive that $p^*_{n,j} = H_{l_n,j}$, which indicates that the design of H guides the optimal prediction p^*. This inspires us to design H so as to estimate the real distribution of $p_{i,j}$, the real probability to classify a sample to the j^{th} class which actually belongs to the i^{th} class as ground truth. We thus calculate the real probability by counting individual AMT users' discrete annotations on defect severity. For a certain defect,

$$\begin{aligned} &H_{i,j} = P\big(annotation = c(j)\big|gt = c(i)\big) \\ &= \textstyle\sum_{x_1,x_2,x_3,x_4,x_5 \in X} P\big(\frac{1}{5}\sum_{k=1}^{5} x_k = c(j)\big|gt = c(i)\big) \\ &= \textstyle\sum_{\frac{1}{5}\sum_{k=1}^{5} x_k = c(j)} P\big(x_1,x_2,x_3,x_4,x_5\big|gt = c(i)\big) \\ &= \textstyle\sum_{\frac{1}{5}\sum_{k=1}^{5} x_k = c(j)} \prod_{k=1}^{5} P\big(x_k\big|gt = c(i)\big) \end{aligned}, \quad (4)$$

where x_k represents each of the five users' annotations, X represents the score set $\{-1.0, -0.5, 0.0, 0.5, 1.0\}$ for *over/under saturation* and $\{0.0, 0.5, 1.0\}$ for the other defects. Function $c(\cdot)$ maps from class label to the center defect severity score for the class. The last equation holds based on a Naive Bayes assumption that all users' annotations are independent of each other given the ground truth.

$$\begin{aligned} &P\big(x_k\big|gt = c(i)\big) \\ &= \frac{P\big(x_k, gt=c(i)\big)}{P\big(gt=c(i)\big)} = \frac{\#\big(x_k \ \&\ gt=c(i)\big)}{\#\big(gt=c(i)\big)} \end{aligned}, \quad (5)$$

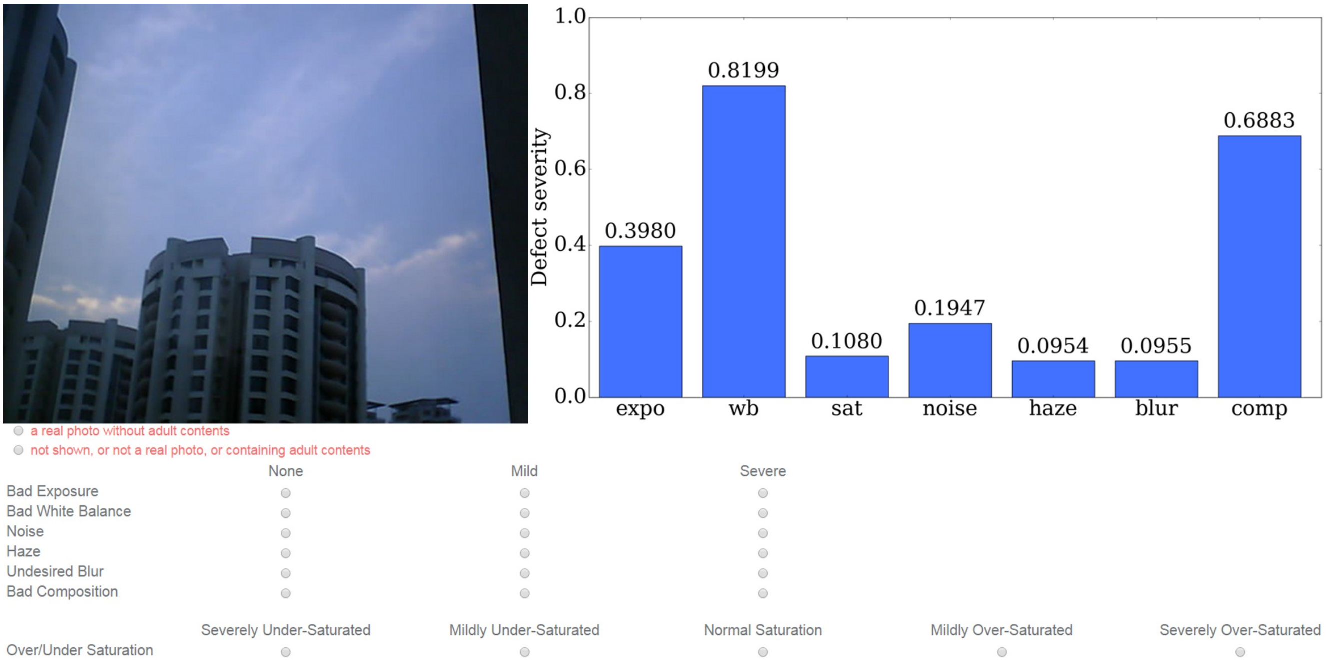

Figure 1. Top left and bottom: One example of our user interface in AMT to collect users' annotations. Top right: for each defect (from left to right: *bad exposure*, *bad white balance*, *over/under saturation*, *noise*, *haze*, *undesired blur*, *bad composition*), we report the weighted average severity ground truth. Higher numbers indicate more severe defects.

which can be directly counted by the frequency from individual users' case-by-case annotations. Based on Equation 4, Figure 4 visualizes our design of infogain matrix H for each defect.

2.2. Data Augmentation

In general, given a defect severity ground-truth histogram, we attempt to make the final histogram more uniformly distributed after augmentation. We augment samples in inverse proportion to class member counts but clap the minimal number as 5 and the maximal number as 50. The minimal number ensures the representativeness of each sampling while the maximal number avoids heavy overlapping.

For the holistic input, each sample is a holistic image with half of the original height and width randomly cropped at the original resolution, which is then warped and downsized to $224 \times 224 \times 3$. For the patch input, each sample is a $224 \times 224 \times 3$ local patch randomly cropped at the original resolution. Additional augmentation by horizontal flipping follows after sampling. We consistently assign all augmented samples with the same severity ground truth as the original image.

Note that there is no augmentation sampling for the *bad composition* defect because image composition is sensitive to the cropping operation.

Figure 5 shows the severity distributions before and after augmentation for each defect. We conclude that our data augmentation has a beneficial rebalancing effect.

3. Experiments

3.1. Evaluation on Synthetic Data

We briefly explain how we generated each synthetic defect. For the exposure defect, we multiplied the intensity by 11 gains. The under-exposure gains have logarithm uniformly spaced in the range $[-1.0, 0.0]$, while over-exposure gains have logarithm uniformly spaced in $[0.0, 1.0]$. For the saturation defect, we scaled the difference between the color and greyscale image by 21 gains with logarithm uniformly spaced in $[-1.0, 1.0]$. For the noise defect, we added white Gaussian noise: 11 Gaussian σ values are uniformly spaced in $[1/255, 22/255]$. For the motion blur defect, we convolve with 11 diagonal blur kernels formed by normalizing the first 11 identity matrices. Each synthetic adjustment is applied to between 420 and 940 testing images labeled as defect-free (the absolute value of severity ground truth smaller than 0.05).

3.2. More Detection Results

Figure 6 and 7 visualize more examples of our testing images, the relative rankings of severity ground truth, and the relative rankings of our predictions.

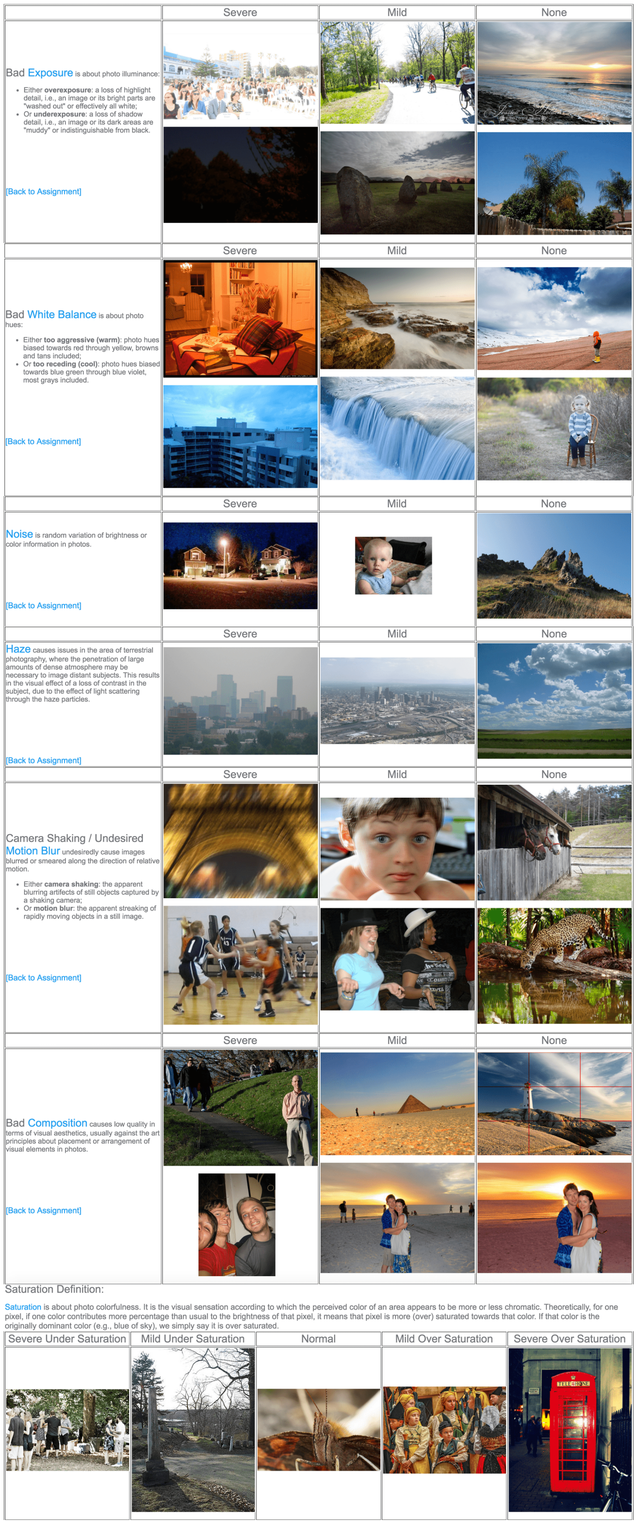

Figure 2. Our instruction Web page that educates AMT users with definitions from Wikipedia and examples for each defect.

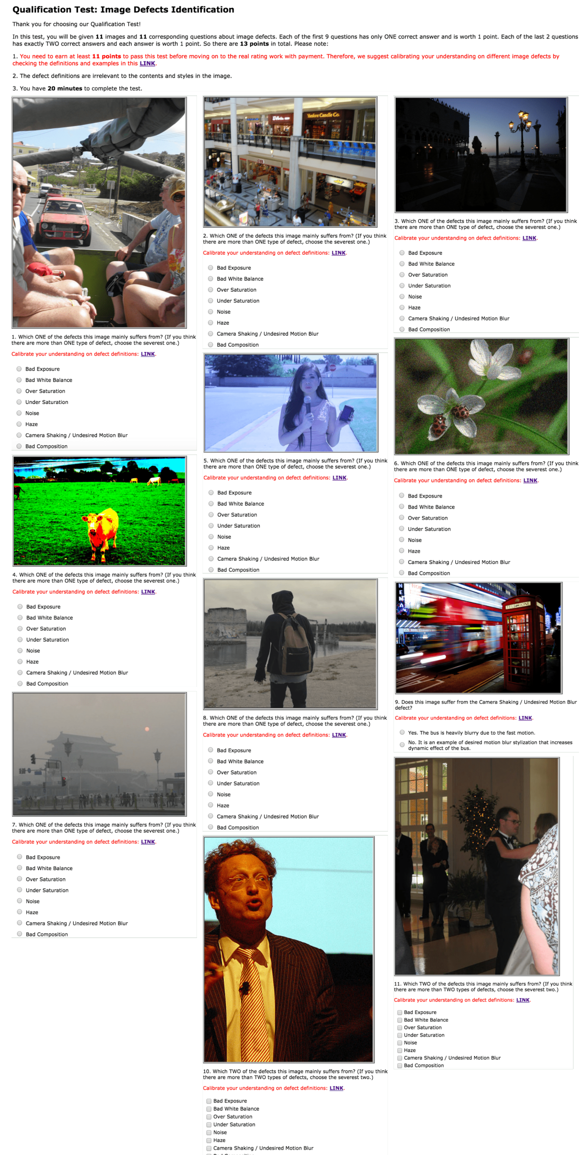

Figure 3. Our educational questions in the qualification test.

3.3. More Localization Results

Figure 8-13 demonstrate more examples of the spatial score maps for different defects.

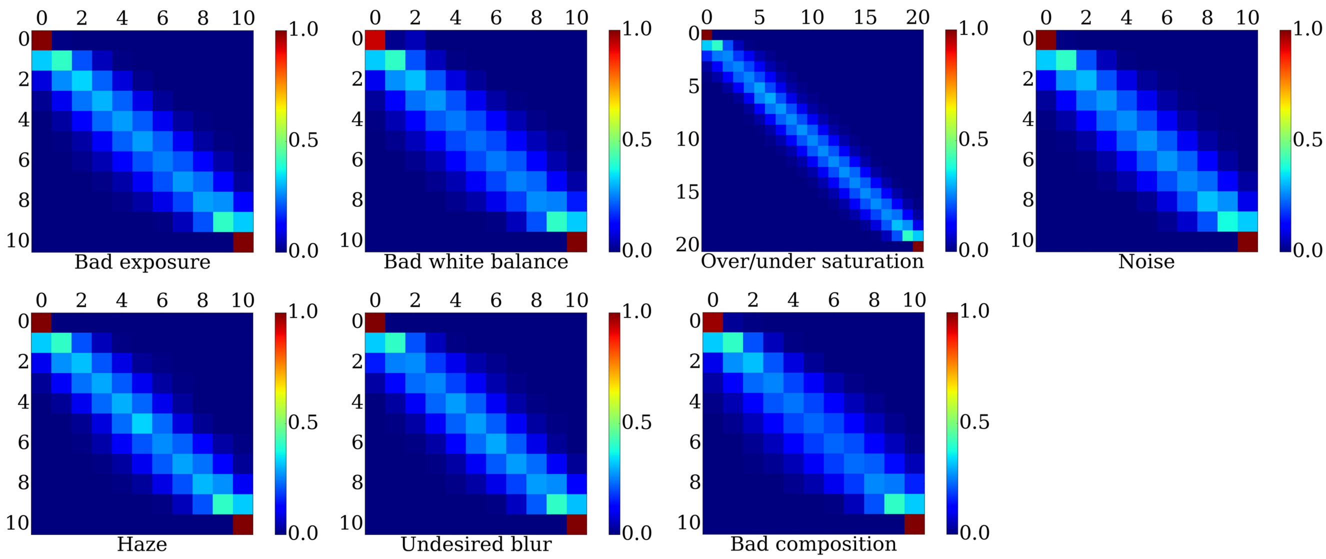

Figure 4. Visualization for our defect-specific infogain matrices. From top left to bottom right: *bad exposure*, *bad white balance*, *over/under saturation*, *noise*, *haze*, *undesired blur*, and *bad composition*. Each row in a matrix corresponds to a class ground truth, and each column corresponds to a classifier prediction. Note that the matrices are asymmetric. The first and last rows represent all users being in agreement that the ground truth is defect-free, or severely defective, respectively. The energy in the first and last rows can be interpreted as strongly encouraging the classifier to perform the same as humans when all humans are in agreement.

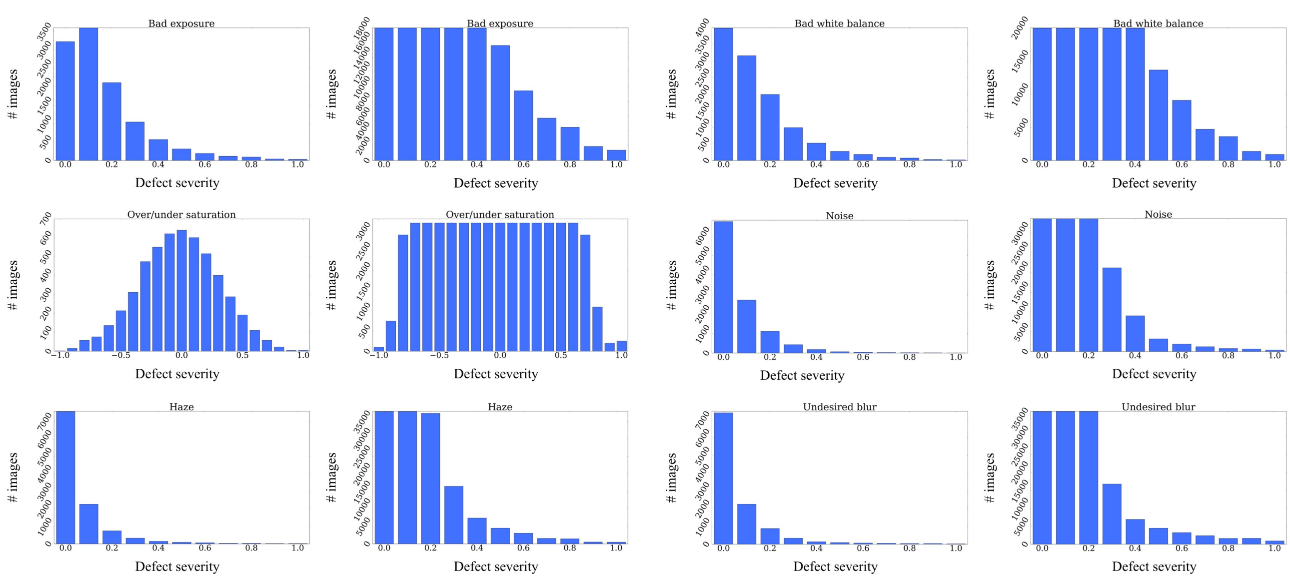

Figure 5. A pair of severity distributions before (left) and after (right) data augmentation for each defect.

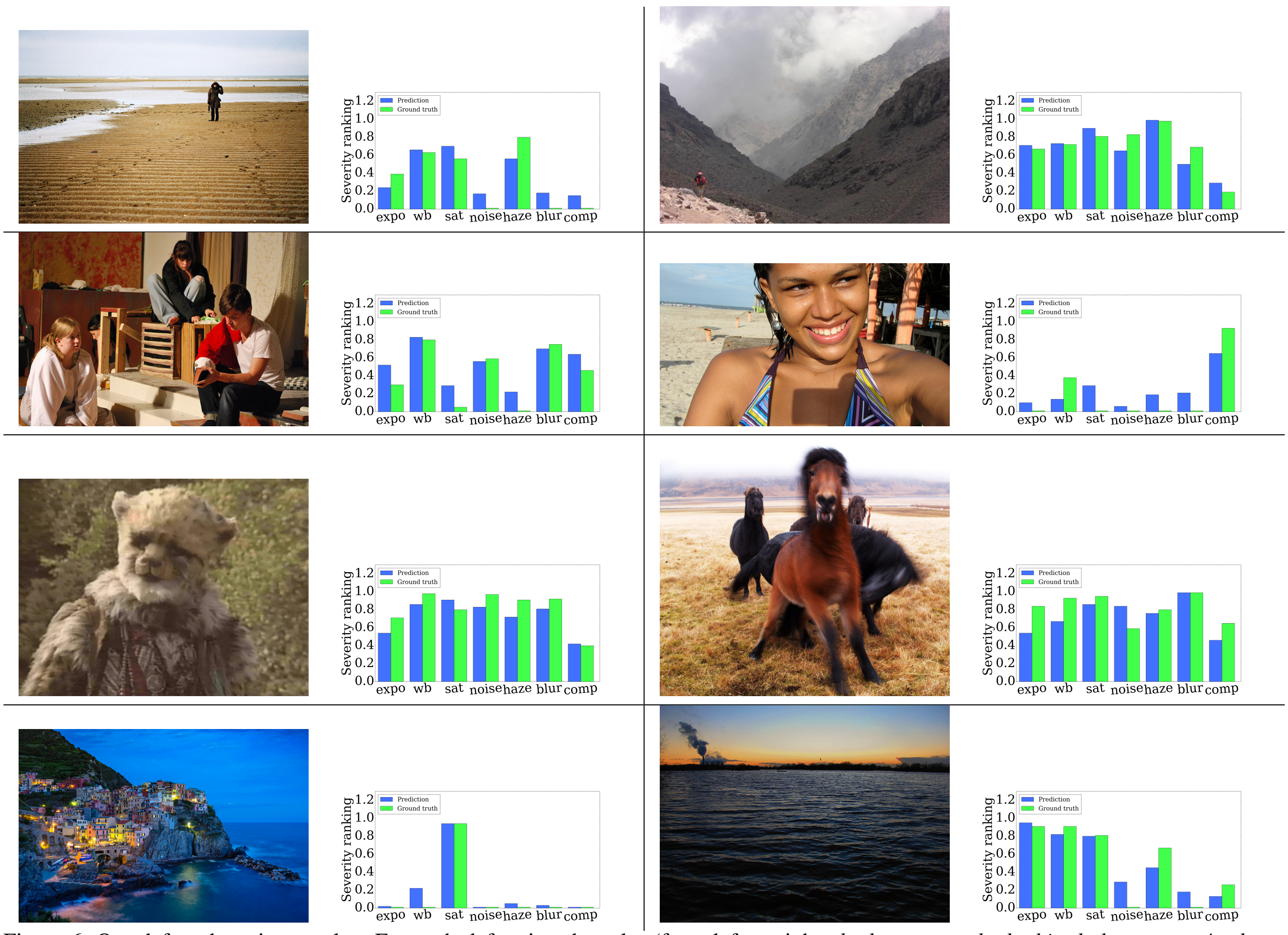

Figure 6. Our defect detection results. For each defect in a bar plot (from left to right: *bad exposure*, *bad white balance*, *over/under saturation*, *noise*, *haze*, *undesired blur*, *bad composition*), we report the relative ranking of a severity score in percentage, which measures the defect severity of a given image compared to all the other photos in a testing set. Higher numbers indicate more severe defects. Our prediction rankings (blue) are consistent with the human judgment (green).

Figure 7. Our defect detection results. For each defect in a bar plot (from left to right: *bad exposure*, *bad white balance*, *over/under saturation*, *noise*, *haze*, *undesired blur*, *bad composition*), we report the relative ranking of a severity score in percentage, which measures the defect severity of a given image compared to all the other photos in a testing set. Higher numbers indicate more severe defects. Our prediction rankings (blue) are consistent with the human judgment (green).

Figure 8. Examples of *bad exposure* defect localization, where the amount of red color indicates the severity of defects in a local region.

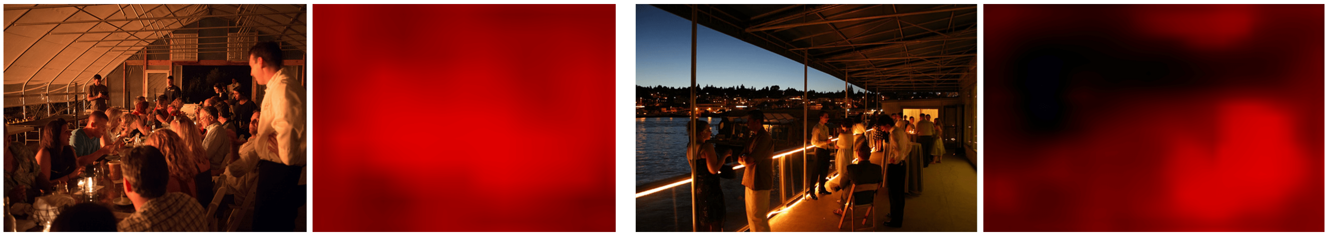

Figure 9. Examples of *bad white balance* defect localization, where the amount of red color indicates the severity of defects in a local region.

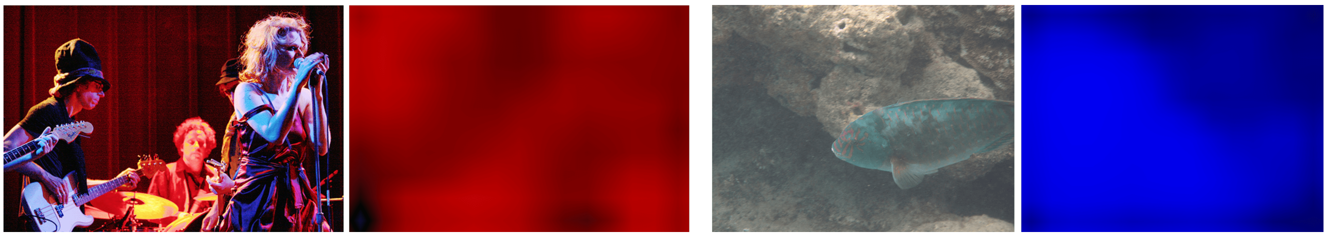

Figure 10. Examples of *over/under saturation* defect localization, where the amount of red/blue color indicates the severity of over/under-saturation defects in a local region.

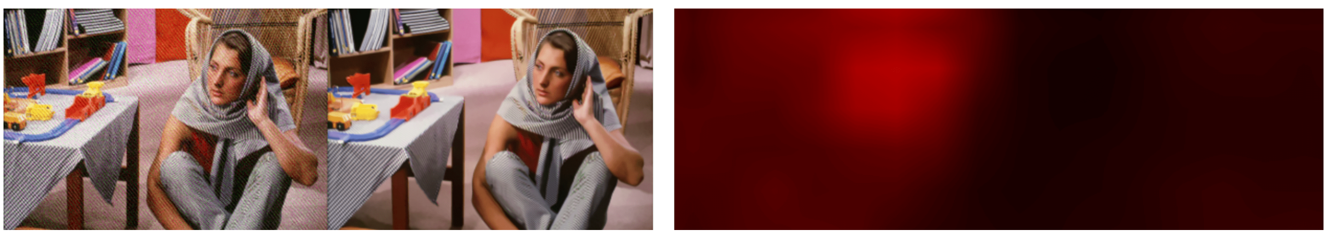

Figure 11. Examples of *noise* defect localization, where the amount of red color indicates the severity of defects in a local region.

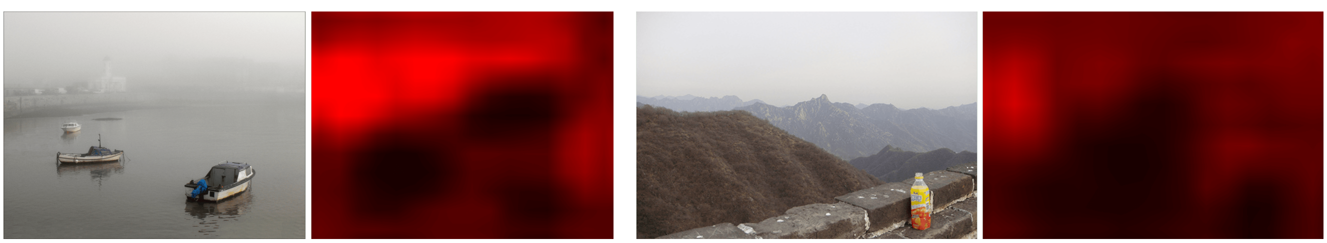

Figure 12. Examples of *haze* defect localization, where the amount of red color indicates the severity of defects in a local region.

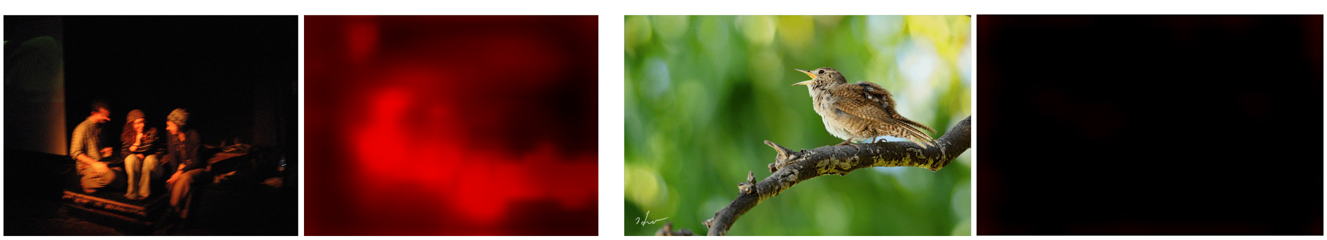

Figure 13. Examples of *undesired blur* defect localization, where the amount of red color indicates the severity of defects in a local region. This pair shows our model has the ability to differentiate between undesired blur and desired depth-of-field effect.