9-1-2019

Assessing neural network scene classification from degraded images

Timothy Tadros
University of California, San Diego

Nicholas C. Cullen
University of Pennsylvania Perelman School of Medicine

Michelle R. Greene
Bates College, mgreene2@bates.edu

Emily A. Cooper
Berkeley School of Optometry

Recommended Citation

Tadros, T., Cullen, N. C., Greene, M. R., & Cooper, E. A. (2019). Assessing neural network scene classification from degraded images. ACM Transactions on Applied Perception, 16(4) https://doi.org/10.1145/3342349

Assessing Neural Network Scene Classification from Degraded Images

TIMOTHY TADROS, University of California, San Diego, USA
NICHOLAS C. CULLEN, University of Pennsylvania, USA
MICHELLE R. GREENE, Bates College, USA
EMILY A. COOPER, University of California, Berkeley, USA

Scene recognition is an essential component of both machine and biological vision. Recent advances in computer vision using deep convolutional neural networks (CNNs) have demonstrated impressive sophistication in scene recognition, through training on large datasets of labeled scene images (Zhou et al. 2018, 2014). One criticism of CNN-based approaches is that performance may not generalize well beyond the training image set (Torralba and Efros 2011), and may be hampered by minor image modifications, which in some cases are barely perceptible to the human eye (Goodfellow et al. 2015; Szegedy et al. 2013). While these "adversarial examples" may be unlikely in natural contexts, during many real-world visual tasks scene information can be degraded or limited due to defocus blur, camera motion, sensor noise, or occluding objects. Here, we quantify the impact of several image degradations (some common, and some more exotic) on indoor/outdoor scene classification using CNNs. For comparison, we use human observers as a benchmark, and also evaluate performance against classifiers using limited, manually selected descriptors. While the CNNs outperformed the other classifiers and rivaled human accuracy for intact images, our results show that their classification accuracy is more affected by image degradations than human observers. On a practical level, however, accuracy of the CNNs remained well above chance for a wide range of image manipulations that disrupted both local and global image statistics. We also examine the level of image-by-image agreement with human observers, and find that the CNNs' agreement with observers varied as a function of the nature of image manipulation. In many cases, this agreement was not substantially different from the level one would expect to observe for two independent classifiers. Together, these results suggest that CNN-based scene classification techniques are relatively robust to several image degradations. However, the pattern of classifications obtained for ambiguous images does not appear to closely reflect the strategies employed by human observers.

CCS Concepts: • **Computing methodologies → Scene understanding**; **Image processing**; *Supervised learning by classification*; *Neural networks*; • **Applied computing → Psychology**;

Additional Key Words and Phrases: Human perception, human scene recognition

EAC was supported by research gifts from Oculus and Microsoft.

Authors' addresses: T. Tadros, 9500 Gilman Drive, Neurosciences Graduate Program, University of California, San Diego, La Jolla, CA 92092; email: tttadros@ucsd.edu; N. C. Cullen, John Morgan Building - Room 132, University of Pennsylvania, Perelman School of Medicine, Philadelphia, PA 19104; email: ncullen@seas.upenn.edu; M. C. Greene, 3 Andrews Rd., Hathorn Hall, Bates College, Program in Neuroscience, Lewiston, ME 04240; email: mgreene2@bates.edu; E. A. Cooper, Minor Hall, University of California, Berkeley, School of Optometry, Helen Wills Neuroscience Institute, Berkeley, CA 94720-2020; email: emilycooper@berkeley.edu.

1544-3558/2019/09-ART21
https://doi.org/10.1145/3342349

ACM Reference format:
Timothy Tadros, Nicholas C. Cullen, Michelle R. Greene, and Emily A. Cooper. 2019. Assessing Neural Network Scene Classification from Degraded Images. *ACM Trans. Appl. Percept.* 16, 4, Article 21 (September 2019), 20 pages.
https://doi.org/10.1145/3342349

1 INTRODUCTION

Recognizing the type of scene depicted in an image or video provides key contextual information with which other visual content—such as objects, actions, and people—can be disambiguated, recognized, and interpreted (Greene and Oliva 2009b; Groen et al. 2017). Humans perform scene classification with surprisingly limited spatio-temporal resolution (e.g., with brief exposure times, in the far periphery, using small image thumbnails), suggesting that at least some aspects of this task can be performed with visual information and analysis of limited complexity (Boucart et al. 2013; Greene and Oliva 2009a; Torralba 2009). In computer vision, many different approaches to scene recognition have been implemented, relying on visual features ranging in complexity from color histograms to spatial patterns to the occurrence of characteristic object classes (Greene 2013; Lazebnik et al. 2006; Li et al. 2010; Martin et al. 2001; Oliva and Torralba 2001; Renninger and Malik 2004; Vogel et al. 2007; Xiao et al. 2010). However, none of these approaches yield fully automatic scene recognition that rivals human levels of performance.

Recent advances in computer vision have led to rising popularity for using deep convolutional neural networks (CNNs) to perform a range of visual tasks. Based on earlier neural networks that were explicitly inspired by biological visual systems (Fukushima 1988), current CNNs show striking performance in visual recognition, due at least in part to the large amount of image data at hand with which to train them (LeCun et al. 2015). In fact, CNN systems trained to perform object recognition now achieve classification rates that rival those of human observers (Russakovsky et al. 2015). In the area of scene recognition, a recent report described a CNN able to classify 365 categories of scenes with a top-five accuracy of over 85% (Zhou et al. 2018).

To achieve this level of accuracy, CNNs are typically trained with images of scenes that are available to researchers on the internet. Such images reflect an intentional process on the part of the photographer to depict the content of the environment in the most informative or aesthetically pleasing manner (Tatler 2007). In the real world, scene or contextual information may not always be the focus of image capture, and thus may be occluded, blurred, or otherwise distorted (Peterson et al. 2016). While human observers show extensive tolerance to such distortions, we do not yet know the extent to which CNNs share this tolerance (Geirhos et al. 2018).

In this article, we address the question of CNN-based scene classification with degraded images using a standard indoor/outdoor classification task. A detailed understanding of the impact of image quality on scene classification is important both (1) for the potential application of CNN-based scene classifiers for real-world tasks, and (2) for the burgeoning interest in using CNNs as a scientific model for biological vision. Thus, we employ both standard (blurring, adding noise) and more exotic (pixel scrambling, grid overlaying) image manipulations to explore how the disruption of natural image statistics affects scene classification. We compare CNN classifications to the responses of a set of human observers, as well as to two baseline classifiers—one lower bound classifier relying on only three hue-saturation-value descriptors per image, and one classifier using prior state-of-the-art GIST descriptors (Oliva and Torralba 2001). First, we review relevant previous work on the classification of degraded images, and the use of CNNs as a model for biological vision.

2 BACKGROUND

2.1 Classification of Degraded Scene Images

A few previous studies have examined how degradation and manipulation of images affects automated scene classification using non-neural network classifiers. These studies have focused on the impact of a specific manipulation: scrambling blocks of image pixels. While large displacements in pixel locations are unlikely to occur during natural image capture, this manipulation is used to disrupt global image structure that may be particularly useful for scene classification (e.g., a horizon line, buildings, walls of a room). Furthermore, scrambling is known to interfere with both object and scene categorization in humans (Biederman 1972; Biederman et al. 1973; Vogel et al. 2007). Using 14 computational models built for both scene and object classification, one study measured various models' performance on classifying images scrambled with 6×6 pixel blocks (Borji and Itti 2014). The authors found that the best classifiers performed similarly to or better than humans on outdoor images, but most of the classifiers performed worse on indoor images. A related study compared a bag-of-words classifier with people's performance on images with scrambled and missing pixel blocks (Parikh 2011). This model performed similarly to people on an outdoor dataset and worse than people on an indoor dataset, mirroring the results of the larger-scale study. While this prior work provides important insights into the type of image information that may be useful for scene classification (local versus global), these studies did not directly address classifier accuracy with degradations that are likely to occur under real-world conditions, nor did they include the more recently developed CNN approaches.

However, several previous studies have examined how image degradations affect the visual classification performance of CNNs on tasks other than scene recognition. For example, Dodge and Karam (2016) quantified how object classification accuracy decreases when images are degraded by blur, noise, compression, and contrast reduction. They compared several different network architectures and found that all architectures were robust to compression artifacts and lowered-contrast, but showed poor performance with blur and noise. In a follow up study, the authors found that fine-tuning CNNs on degraded images could improve classification accuracy of blurred and noisy images, but performance was still worse than a comparison dataset of human classifications (Dodge and Karam 2019). Using similar manipulations, Vasiljevic et al. (2016) and Zhou et al. (2017) also showed that substantial reductions in object classification accuracy can be rescued by fine-tuning CNNs, but also that performance on semantic segmentation is difficult to fine-tune. In a face-recognition study, Karahan et al. (2016) showed that blur and noise reduce CNN accuracy more so than manipulations of overall contrast and color. In addition, they reported that occluding different regions of the face variably impacted classification, with eye occlusions being particularly problematic. With the exception of Dodge and Karam (2019), the performance changes reported in these studies were not benchmarked against how the manipulations affect human observers, so it is unclear whether the performance reductions are expected based on human perception. Recently, Geirhos et al. (2018) examined the differences between several CNN architectures and the performance of human observers on an object classification task in a tightly controlled psychophysical setup. They found that CNNs outperformed human observers for full-color, full-contrast, un-manipulated images, but that human observers were more robust to the deletion of color, decreases in contrast, or the addition of noise. However, to date, there is no study that compares human observers, CNNs, and other machine vision approaches for scene classification performance.

With respect to noise specifically, it is well documented that CNNs can be vulnerable to adversarial noise: the subtle addition of noise that, while undetectable to human observers, leads a CNN to alter its classification verdict with high confidence (Goodfellow et al. 2015; Szegedy et al. 2013). The relative fragility of CNN performance compared to human observers could suggest that scene classifiers based on pre-defined descriptors, such as GIST, may outperform CNNs when the signal-to-noise ratio gets lower. Thus, in the current article, we include the performance of a GIST-based classifier to examine whether image degradations disproportionately affect the performance of CNNs.

2.2 CNNs as a Model for Biological Vision

Given the similar level of classification performance between CNNs and human observers on intact images, an active line of current research involves assessing the extent of representational similarity between the two (Cadieu et al. 2014; Cichy et al. 2016; Geirhos et al. 2018; Groen et al. 2018; Güçlü and van Gerven 2015; Khaligh-Razavi and Kriegeskorte 2014; Kubilius et al. 2016; Yamins et al. 2013). Specifically, if a CNN is to be used as a model for human vision, it is essential not just that it achieves human-level accuracy, but that the information representation is similar (Khaligh-Razavi and Kriegeskorte 2014; Kriegeskorte et al. 2008; Yamins and DiCarlo 2016). If this representational similarity is high, it has been suggested that researchers can use neural networks to better understand how biological visual processing works (Kriegeskorte 2015). Specifically, having a more similar model of the human visual system will provide insight into the unique failures and successes of visual processing in people and how visual information is used to motivate specific visual tasks.

With respect to direct comparisons between CNNs and biological neural networks, some researchers have shown that appropriately trained CNNs can predict neuronal firing in inferior temporal cortex of non-human primates as well as the overall neuronal firing code (Khaligh-Razavi and Kriegeskorte 2014; Yamins et al. 2013). CNNs have also been shown to predict human brain activity in fMRI (Agrawal et al. 2014; Güçlü and van Gerven 2015), MEG/EEG (Cichy et al. 2016; Greene and Hansen 2018), as well as behavioral judgments (Greene et al. 2016; Jozwik et al. 2017; Kubilius et al. 2016). However, other work suggests that while category-level judgments in object recognition are well modeled by CNNs, these networks are not predictive of performance on an image-by-image basis (Rajalingham et al. 2018). As GIST features have also been shown to explain a sizeable amount of variance in the response patterns of human visual cortex (Watson et al. 2017), we believe that it is useful to directly compare this representation to that of a CNN. Thus, in addition to reporting how image degradations affect the scene recognition accuracy of humans and CNNs, in the current study, we also explore the level of image-by-image agreement on classifications between human observers and our classifiers of interest.

3 METHODS

3.1 Image Selection

We selected 200 images from the Places-205 database (Zhou et al. 2014) to assess the impact of image degradation on human and computer scene classification. Places-205 contains nearly 2.5M images with 205 scene categories. In the current project, we focus on indoor/outdoor scene classification, as an example of a challenging classification task that requires generalizing over highly diverse images. To create our sample, we chose 10 indoor and 10 outdoor subcategories and selected 10 images at random from each of these subcategories (100 indoor and 100 outdoor images total).

For indoor subcategories, we used: bedroom, classroom, dining room, kitchen, living room, lobby, museum, office, restaurant, and supermarket. For outdoor subcategories, we used: coast, forest path, highway, mountain, skyscraper, valley, sea cliff, river, residential neighborhood, and snowfield. We selected these subcategories to capture the variability of indoor and outdoor scenes, including outdoor natural scenes and outdoor man-made scenes, as well as indoor scenes that reflect both public and private spaces (Tversky and Hemenway 1983). All images were 256×256 8-bit RGB format, and were taken from the validation set in the Places-205 dataset. These 200 images were not used in training either the original PlacesNet-205 or the other classifiers used herein.

By selecting 200 images out of 2.5M, we risk selecting a sample that is not representative of the whole set. We attempted to avoid substantial bias using a combination of manual and random selection. To examine how representative the selected 200 images were, we conducted a comparison of their visual statistics based on GIST descriptors (Oliva and Torralba 2001). For the indoor and outdoor test samples separately, we computed the probability distribution of each GIST descriptor and compared this to that descriptor's distribution across all indoor and outdoor images in the Places-205 database using the Kullback-Leibler divergence (512 descriptors total). This statistic provides a measure of how much the distribution of each descriptor differed between the smaller test samples and all images within that category. Next, we iteratively sampled a different 100 indoor or 100 outdoor images from the database at random without replacement and repeated this calculation 1000 times. Across all descriptors, we found that on average our samples' divergence values fell at the 48th percentile (standard deviation = 30) for indoor images and 68th percentile (standard deviation = 29) for outdoor images. These percentiles suggest that the selected sample images are about as representative as would be expected in

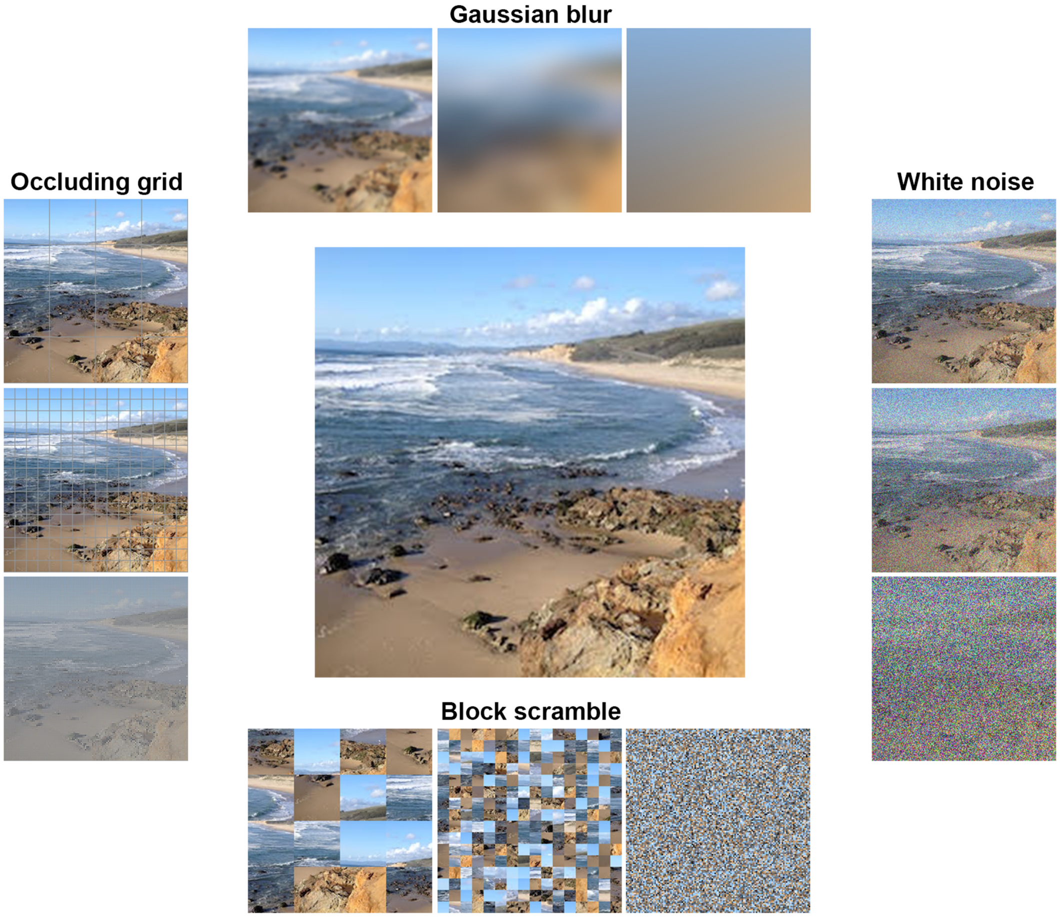

Fig. 1. Example image manipulations. An un-manipulated outdoor image is shown in the center (note this example is not from the Places-205 database). Illustrated blur levels correspond to kernel standard deviations of 2, 16, and 128 pixels. Noise levels correspond to noise weights of 0.2, 0.4 and 0.8 noise. Scramble and occluding grid levels correspond to blocks of width 64, 16, and 2 pixels.

terms of low-level visual statistics had we instead selected images at random from Places-205 database. While the outdoor sample may be less representative of the database by this low-level measure, the selected subcategories also provide a broad sampling of the diversity of outdoor scenes.

3.2 Image Manipulation

We employed a range of different methods to degrade these images, and we varied the intensity of each method from low to high (examples shown in Figure 1):

- *Gaussian blur*: Image detail was removed by convolving with a 2D isotropic Gaussian blur kernel with standard deviations of 2, 4, 8, 16, 32, 64, and 128 pixels. Replication padding was used to reduce edge artifacts.
- *White noise*: A weighted combination of each image was taken with a 256×256×3 white noise pattern. Image weights were varied from 0.8 to 0.2 in four steps. The weights given to the image and the noise pattern always summed to one.

- *Scramble*: Images were segmented into square blocks of width 2, 4, 8, 16, 32, 64, and 128 pixels, and each block was mapped to a new random location in the image. We scrambled images both with and without a grid line marking each block. The results were qualitatively similar, so we focus on the results for scrambling without added grid lines and include the results with grid lines in the Supplementary Material.
- *Occluding grid*: Local image detail was removed by replacing pixels with the average RGB value of the image in a grid pattern. Each line on the grid was 1 pixel wide, and the spacing was 2, 4, 8, 16, 32, 64, or 128 pixels. This created an occluding grid across the entire image.

We also included two color manipulations. We converted the images to grayscale using the equation: 0.2989R + 0.5870G + 0.1140B, where R, G, and B refer to the red, green, and blue color channel values, respectively. We also created color-inverted images, in which pixel values are subtracted from the maximum within each color channel (not shown in Figure 1). After these manipulations, each of the 200 scenes had 35 versions, including the original unmanipulated image, creating a test set of 7K images.

3.3 Human Performance

To assess human scene classification performance on the test images, we recruited participants on Amazon Mechanical Turk. Participants were asked to classify the test set images as indoor or outdoor (single-interval binary choice). Participants viewed the images one at a time, and each participant saw each unique scene once (in random order), with a randomly selected manipulation (200 trials total). This was done so participants could not use information from previous trials to improve their classifications (e.g., a blurry version of an image may be easier to classify if the original was seen on a previous trial). There was no time limit for making responses. Four additional images were included as catch trials. These images were easily identifiable, intact images. Prior to beginning data collection, we set an exclusion criterion of making an incorrect response on any one of these catch trials. We obtained completed datasets from 216 participants and excluded results from 9 individuals based on this criterion. In total, we analyzed results from 207 participants. Due to the random selection of images for each observer, the total number of presentations varied across images. A small percentage of the 7K images (0.3%) were never shown to the participants, and were thus excluded from further analysis. Of the remaining images, the median number of presentations was 5 (minimum of 1 and maximum of 15). The main analyses focus on accuracy and agreement aggregated within a given manipulation, for which the median number of presentations was 1,185 (minimum of 1,126, maximum of 1,233). Participants were compensated for their time, and the study procedures were approved by the institutional review board at Dartmouth College.

3.4 Neural Networks

We used a pre-trained CNN for scene classification: Places-205-AlexNet (abbreviated as PlacesNet), which uses the neural network architecture of AlexNet (Krizhevsky et al. 2012) and was trained on the 205 scene categories in the Places-205 database (Zhou et al. 2018). The vast majority of images in the Places-205 database are 8-bit jpeg images with three-channel RGB color. Before being input to the network, the 256×256 images are down-sampled to 227×227. The input is then fed through a series of five convolution, pooling, and normalization layers. The pooling layers implement maximum pooling and downscale the output size to reduce the number of parameters. The normalization layers refactor the input features such that the inputs into each non-linearity are zero-centered and have unit-variance. This ensures that each epoch of training operates on features with the same distribution and statistics. Finally, there are three fully connected layers; the last of which produces a 205-dimensional vector containing the softmax probabilities for the 205 classes in the Places-205 database. We used a majority vote for the top-five categorization to determine the indoor or outdoor class of the image. Specifically, we took the five most-likely scene classes and counted how many were outdoor or indoor. The classification label was then taken to be the majority class (outdoor/indoor) in the top-five.

We also created a modified version of PlacesNet trained directly to classify images as indoor or outdoor (InOutNet). This network architecture was identical to PlacesNet, with the exception that the size of each of the layers was reduced by a factor of two (excluding the first convolutional layer), and the final fully connected layer contained only two units, reflecting indoor and outdoor categories. We trained this network on a subset of the Places dataset, using the indoor/outdoor category labels only. The subset was selected to have a matched number of indoor and outdoor images, and so all images contained three-channel RGB color (some images in the Places database are grayscale). This resulted in 710,052 training images in each category. For this network, the classification label was taken to be the class with the highest score.

Training and testing were implemented through the Caffe neural network library (Jia et al. 2014). The InOutNet was trained with the same parameters used to train PlacesNet-205-AlexNet (Zhou et al. 2016). A softmax loss layer was added to the neural network to compute the learning objective. Training was performed for a predetermined number of iterations (300,000) with batch sizes of 256. Before training, InOutNet had an accuracy of 33.7% on the validation set and a loss of 0.697. After 300,000 iterations of training, the network improved to 93.86% classification accuracy with a loss of 0.169. These loss values are computed using the softmax function and represent how well the learning objective is achieved.

3.5 LDA-based Classifiers

In addition to comparing the CNN scene classification performance to human observers, we also wanted to examine the performance of these classifiers relative to previously proposed scene recognition techniques. This serves both as a second benchmark, as well as a reasonable lower bound on performance.

First, we computed the GIST image descriptors described in Oliva and Torralba (2001) and used extensively in the scene recognition literature. These descriptors represent relatively global properties of a scene's spatial structure. Using four spatial scales, eight orientations, and segmentation into 16 non-overlapping regions (each region is 64×64 pixels), we computed 512 descriptors for each image in the balanced indoor/outdoor image set used to train InOutNet. We used linear discriminant analysis to compute a linear decision boundary between the indoor and outdoor image classes. We refer to this classifier as GIST-LDA. To compute a lower bound on expected performance, we also trained a simple linear classifier using just three descriptors from each image—the mean hue, saturation, and value (HSV-LDA).

3.6 Analysis of Accuracy

For the four classifiers (PlacesNet, InOutNet, GIST-LDA, HSV-LDA), accuracy was computed in terms of percent correct over all images. For the human observers, each image had more than one response. Thus, we computed the number of times each image was presented across all participants and the number of times a correct classification was given. These values were summed over all images to determine percent correct. Confidence bounds (95%) for the accuracy of both the humans and the classifiers for each image manipulation were computed using the binomial distribution.

To examine whether the performance of each classifier was significantly better or worse than the human observers, we computed the observed accuracy ratio ($\hat{r}$) between the classifier and the human observers for each manipulation as

$$\hat{r} = \frac{k_c/n_c}{k_h/n_h}, \tag{1}$$

where k_c and k_h denote the number of correct responses obtained from the classifier and human participants, respectively; n_c and n_h denote the total number of trials for the classifier and all human participants, respectively. For the classifiers, n_c is always equal to the number of images being analyzed, since there is only one classification per image. We consider $\hat{r}$ as an estimate of the true underlying accuracy ratio (r), with values of r greater than one corresponding to more accurate classifier performance compared to humans for a given sample of images, and values less than one corresponding to less accurate performance compared to humans.

For a given r, we can compute the expected standard deviation of the distribution of $\hat{r}$ given a finite number of trials using the log-method described in Katz et al. (1978) and Koopman (1984):

$$\sigma_{\hat{r}} = \sqrt{\frac{1}{k_c} - \frac{1}{n_c} + \frac{1}{k_h} - \frac{1}{n_h}}. \tag{2}$$

Using this distribution, we determined the p-value associated with $r = 1$ (i.e., the probability of the measured $\hat{r}$ for a given sample of images, if there was actually equivalent performance between classifier and observers), as follows:

$$p = -2\left(\Phi\left(\frac{\pm \ln(1/\hat{r})}{\sigma_{\hat{r}}}\right) - 1\right), \tag{3}$$

where Φ denotes the cumulative of the normal distribution, and the negative natural log is used when $\hat{r}$ is less than 1 (Koopman 1984). In each analysis, p-values were used to determine statistical significance, corrected for multiple comparisons with a false discovery rate of 0.05, as described in Benjamini and Hochberg (1995).

For completeness, we also include a supplementary analysis of sensitivity and bias (Stanislaw and Todorov 1999). We calculated the sensitivity index (d') and criterion (c) as

$$d' = \Phi^{-1}(\text{hit}) - \Phi^{-1}(\text{false}), \tag{4}$$

$$c = \frac{-\left(\Phi^{-1}(\text{hit}) + \Phi^{-1}(\text{false})\right)}{2}, \tag{5}$$

where Φ^{-1} denotes the inverse of the cumulative of the normal distribution, and "hit" and "false" denote the hit rate and false-alarm rate, respectively. By convention, we treated correct classification of indoor images as hits. This criterion provides a measure of response bias, with negative values indicating a bias to classify images as indoors and positive values indicating a bias to classify images as outdoors.

3.7 Analysis of Agreement with Humans

We also wanted to determine whether each classifier tended to agree or disagree with the human participants more than expected by chance. For this analysis, first we computed the number of trials in which human observers gave the same response as the classifier (i.e., the number of agreements). Next, we computed the expected number of agreements, assuming the classifier categorization of individual images as indoor/outdoor was independent from the participant categorization. Intuitively, two independent classifiers with near 100% or near 0% accuracy would *a priori* have to agree on almost all images. Thus, high levels of agreement in these cases are expected. We defined the observed number of trials on which the human participants agreed with the classifiers as a, and the expected number of agreements ($E[a]$) was calculated as:

$$E[a] = n_h\left(\left(\frac{k_c}{n_c}\right)\left(\frac{k_h}{n_h}\right) + \left(\frac{w_c}{n_c}\right)\left(\frac{w_h}{n_h}\right)\right), \tag{6}$$

where w_c and w_h indicate the number of incorrect responses obtained from the classifier and human participants, respectively. Intuitively, this is the total number of trials, multiplied by expected proportion of agreement. The expected proportion of agreement is given by the product of the human/classifier proportions correct added to the product of the human/classifier proportions incorrect. This calculation is similar to measures of inter-rater reliability, such as Cohen's Kappa.

We computed agreement ratios, confidence intervals, and p-values as described for accuracy ratios (Equations 1, 2, 3), reflecting how much the observed proportion of agreement (a/n_h) either exceeded or fell below expectation ($E[a]/n_h$), and whether this deviation differed significantly from chance. For the results shown in Figure 4, we computed this agreement ratio for each classifier across all 7,000 images; for Figure 5, we computed this ratio for images within each manipulation type and level separately. To determine the upper bound on these agreement ratios (that is, the maximum possible agreement ratio for the classifier, given human accuracy), we computed the agreement ratio for 100% agreement (i.e., $1/(E[a]/n_h)$).

It is important to consider that the level of agreement among human observers places an upper bound on the extent of human-classifier agreement. By way of example, if humans tended to classify blurry images correctly only 50% of the time, they could be consistently making the same mistakes (e.g., they tend to classify a blurry image of an office with green furniture as a forest), or each observer could be randomly guessing. If mistakes are consistent across observers, a classifier that reflects human perceptual processes should make the same mistakes. If the mistakes are inconsistent across observers, a classifier should not be reasonably expected to mirror them. To examine the level of agreement between humans, we created an additional consensus classification for each image reflecting the response given by the majority of human observers. We then computed the expected, observed, and maximum possible agreement levels between the individual observers and this consensus. Intuitively, if human observers provided a heterogeneous set of responses for the same images, the agreement ratio with the consensus will be low. If they provided the same response, this ratio will be high. Importantly, the calculated levels of expected agreement and maximum possible agreement are always tied to the baseline accuracy (Equation 6), so it is not possible to directly apply the agreement ratio calculated for the human consensus as an upper bound to the classifiers. To do so, we would have to create conditions under which the humans and classifiers had equivalent accuracy. Nonetheless, this provides a useful piece of information when interpreting the agreement between classifiers and humans.

Lastly, to examine patterns of agreement and disagreement with human responses in more detail, we also conducted an exploratory analysis on the classification "confidence." For the LDA-based classifiers, we used the posterior probability of each class as a measure of confidence. For the InOutNet, we used the output of the softmax function of the final layer. Note that softmax layer activations may not be well-calibrated measures of confidence (Guo et al. 2017); however, for the current analysis, we loosely interpret these values as reflecting class probability. For each individual image, we compared classifier confidence to the probability of indoor/outdoor responses across the human observers. We restricted this analysis to images that had responses from at least four observers (a total of 5,891 out of 7,000).

4 RESULTS

4.1 Classification Accuracy

As expected, the human observers had near-perfect indoor/outdoor classification accuracy on the original, undegraded 200 images (Table 1). Both CNNs performed comparably to the observers, whereas the two classifiers based on GIST and HSV descriptors were substantially less accurate (88.0% and 69.0%, respectively). Averaging over all 7,000 images, which include all degradation levels, human performance dropped by ~12 percentage points, but the CNN and GIST-LDA classifiers' accuracy dropped substantially more (over 20 points each). This suggests that the degradation of the test images hampered the computer-vision classifiers more than human observers. The HSV-LDA classifier's accuracy was not substantially different between the original and degraded images. Intuitively, this is likely because most of the manipulations altered the statistical features of the images captured by the neural networks and GIST, but did not largely affect the global image hue, saturation, and value (except the grayscale and color inversions).

We next examined how different image degradations impacted performance. Figure 2 shows the classification accuracy of the human observers and each of the classifiers for the Gaussian blur, white noise, block scramble, and occluding grid manipulations. The left-most data points in each plot always correspond to the results for the original, un-degraded images. Degradation severity increases from left to right. We first focus on comparing the CNN-based classifiers to the human benchmark. For direct comparison, Figure 3 plots the accuracy of each classifier as a ratio to the human observers' accuracy.

Table 1. Summary of Classification Accuracy

	Original images		**All Images**	
	accuracy	*95% CI*	*accuracy*	*95% CI*
Humans	99.3	(98.6,99.7)	87.0	(86.7,87.3)
PlacesNet	98.5	(95.7,99.7)	74.1	(73.1,75.2)
InOutNet	98.0	(94.9,99.4)	76.3	(75.3,77.3)
GIST-LDA	88.0	(82.7,92.1)	63.5	(62.4,64.6)
HSV-LDA	69.0	(62.1,75.3)	67.6	(66.5,68.7)

Percent correct and 95% binomial confidence intervals (CI) for only the intact originals (first and second columns) and across all images (third and fourth columns).

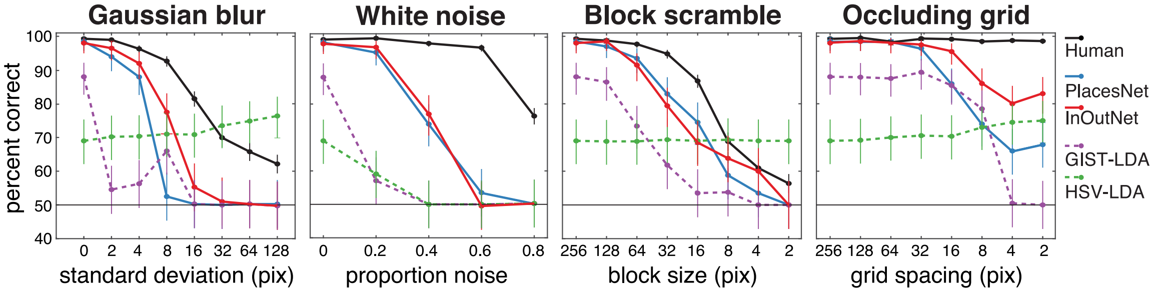

Fig. 2. Classification accuracy as a function of each image manipulation. Error bars represent 95% confidence intervals based on the binomial distribution. Bars are smaller for the human observers, because across observers, more than one trial could be obtained for each image.

For the human observers (Figure 2; black lines), we found that only blur, noise, and scrambling substantially and systematically reduced the ability to classify the scene images. These manipulations also substantially reduced the accuracy of the two CNN-based classifiers (PlacesNet and InOutNet), both in an absolute sense (Figure 2) and in terms of their performance relative to humans (Figure 3). Despite near-human accuracy on the original images, these two neural networks systematically under-performed relative to the human observers as the image degradations increased (red and blue solid lines). The decrease in CNN performance can be seen as a general tendency for the accuracy ratio to dip below 1 in the left two columns of Figure 3. This dip is pronounced for the two manipulations that most disrupted local image features (blur and noise), whereas the CNNs' performance was relatively more robust to the scramble manipulation. This may reflect a certain amount of spatial invariance in the features identified by the neural networks. Recall that the InOutNet was trained specifically for the indoor/outdoor classification task. For some manipulations, this network descriptively outperformed the original PlacesNet on this task (e.g., mid-level blur), although overall the two networks had highly similar results.

Interestingly, the occluding grid manipulation had essentially no impact on human observers (Figure 2, rightmost panel). As can be seen in the example images in Figure 1, even with every other pixel occluded by the grid, the mountain scene is still highly recognizable with human vision. This was not the case for the CNN-based classifiers—once the grid sizes approached the size of the filters in the initial convolutional layer (11×11 pixels), accuracy dropped. Although accuracy on these images stayed well above chance, it was markedly worse than human accuracy for mid- and small-scale grids with both CNN-based classifiers (Figure 3, bottom row).

In comparison to the CNNs, the GIST-LDA classifier is limited to relatively global image descriptors. As expected, this classifier tended to under-perform relative to human observers and relative to the CNNs on the original images (Table 1). One might expect the GIST-based classifier to outperform the CNN-based classifiers on degraded images, because neither CNN was trained on degraded images. However, this is not what we observed

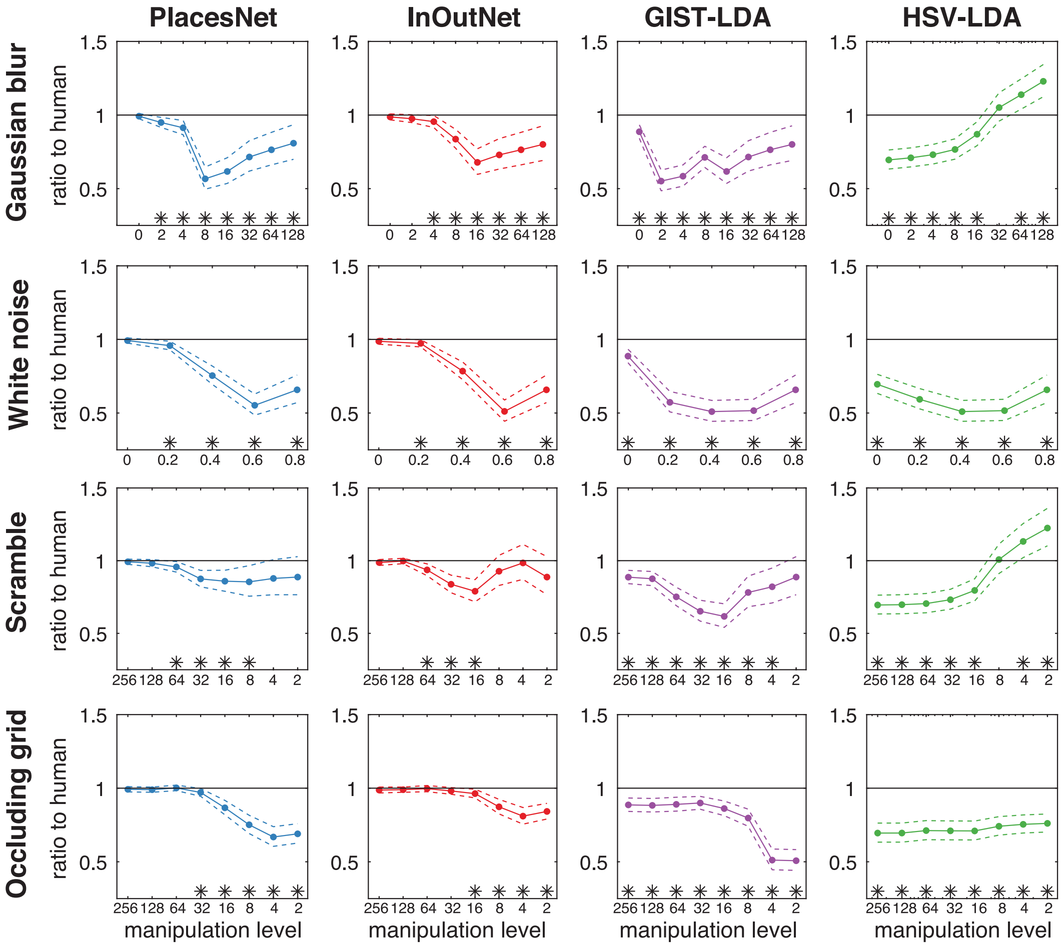

Fig. 3. Classification accuracy relative to human performance. Data points indicate the accuracy ratio of classifiers to human observers for images within each manipulation level. Error bands represent the 95% confidence intervals for this ratio, with ratios significantly different from one indicated with asterisks. Data from the block-scramble-with-grid, as well as color manipulations, are not shown, but these data were included in the FDR correction. Units for manipulation levels (abscissa) are the same as in Figure 2.

(purple lines; Figures 2). The observation that the CNN-based classifiers tended to outperform the GIST-based classifier on degraded images, despite only being trained on the originals, provides a good example of generalization in CNN-based scene classifiers. As shown in Figure 3 (third column), the GIST-based classifier was also less accurate overall than humans on degraded images.

It is notable that our lower-bound HSV-LDA classifier—while worse at most degradation levels—out-performed the other techniques, including humans, for some of the most degraded images (maximum blur and scramble) (green lines; Figures 2 and 3). This makes reasonable sense: By only relying on three completely global image features, this classifier could not perform nearly as accurately as humans on intact images, but was also unaffected by the loss of spatial detail or local structure. The cases in which the HSV-LDA classifier substantially out-performed the CNNs (such as large amounts of blur) suggest that that there is highly global information relevant to scene classification that is not captured by the CNNs. This difference might occur, for example, if the training process for the CNNs places less weight on global information because it has less utility when other information is available.

We also included a block-scramble-with-grid manipulation and two color manipulations (see Methods). Plots of the accuracy results from these conditions are included in the Supplementary Material (Figure S1). As expected, the block-scramble-with-grid manipulation effects were similar to the block scramble. Making the images grayscale and color-inverted only strongly reduced the HSV-LDA performance (relative to accuracy on intact images); however, color inversion also had a moderate negative impact on the CNNs.

Taken together, these results suggest a complex interaction between image degradation and scene classification performance. Not surprisingly, the CNNs performed worse on degraded images, relative to intact originals. However, these classifiers maintained above-chance performance for a wide array of manipulation types and severities. Benchmarked against humans, these networks seem to be more impacted by manipulations that disrupted local image patterns, even when human accuracy was unaffected (e.g., adding a grid). Nonetheless, within a reasonable range of degraded image information, CNNs for scene recognition provide state-of-the-art classifications. In the Supplementary Material (Figures S2 and S3), we include an additional analysis of human and classifier performance in terms of d' and criterion, rather than percent correct. The pattern of results for d' closely follow the percent correct data. The criterion results suggest that all four classifiers become substantially biased under certain conditions, while humans do not. We explore this observation further in the following section.

4.2 Classification Agreement

Strong agreement provides stronger evidence for similar representations across human observers and computer-vision classifiers. This is a necessary precondition for using computer vision as a model for understanding human visual processing. To examine agreement, first the expected value of agreement between each classifier and the human observers was calculated, assuming that the human and computer vision classifications were independent from one another (see Methods). The left panel of Figure 4 shows how each classifier compared to this expected value, as a ratio averaged over all image manipulations. Values greater than 1 indicate that the classifier agreed with human observers more than expected by chance; values less than 1 indicate below chance levels of agreement (i.e., systematic disagreement). For comparison, the agreement ratio for a consensus classification averaged across all human observers is also included (gray bar), which indicates how much individual human observers tended to agree with the average response. It is important to interpret all of these agreement ratios values in light of the maximum possible ratio (ceiling), which depends on the overall accuracy of classification. These ceiling values are also plotted, as the shaded regions above each bar. For the CNN-based and GIST-based classifiers, the observed agreement significantly exceeded chance levels, but remained substantially below ceiling. This result suggests that, at least across these artificially degraded images, the CNN-based classifiers did classify scenes in a way that agreed to some extent with the classification of human observers, but this agreement was not strong. This agreement was descriptively but not substantially more than the GIST-based classifier. Not surprisingly, the HSV-based classifier did not agree with human classifications any more than expected by chance. By comparison, human observers tended to agree with each other at a slightly higher ratio than the classifiers (13% above chance) despite having a lower ceiling. The disagreement between human observers should be taken into consideration, because it necessarily limits the probability that any single classifier can capture the pattern of human responses. To examine whether human observer disagreement was a substantial contributor to the low classifier-agreement ratios, we recalculated the classifier-agreement values using the consensus, rather than the individual observer responses (Figure 4, right panel); as expected, this slightly increased agreement, but did not substantially change the overall pattern of results.

To examine whether different manipulations resulted in classifications that were more or less similar to human observers, in Figure 5, we plot the agreement ratio considering each manipulation type separately. The black line in each panel indicates the ceiling for the agreement ratio at each degradation level. Intuitively, the ceiling tends to be near 1 when the manipulations are weak, because in this case the expected agreement is already very high

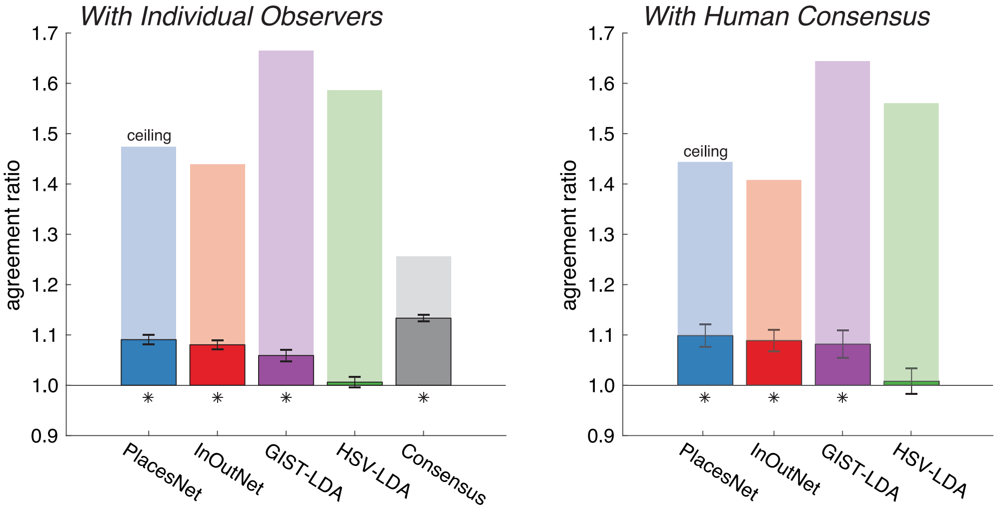

Fig. 4. Bar heights and error bars indicate the measured agreement ratios and 95% confidence intervals. Ratios significantly different from one are indicated with asterisks. Consensus refers to a classification of each image based on the most frequent response across human observers. The left panel shows the results when agreement is calculated across all individual observers; the right panel shows the same analysis when individual responses were replaced by the Consensus. The shaded area above each bar indicates the ceiling, given the accuracy of the classifier.

(Equation 6). Thus, we cannot use these conditions to conclude one way or another about how much systematic agreement we observe. For the blur and noise manipulations, when the ceiling increases for higher levels of degradation, there is a tendency for the agreement ratio for each classifier to increase slightly as well. For the scramble manipulation, the results are more mixed, with some conditions leading to systematic disagreement. Much like with the accuracy analysis, these data show that the effect of the occluding grid (bottom row) on human classifications was poorly modeled by any of the classifiers. This result suggests a substantial difference in the way that CNNs and humans utilize mid-scale image features. In particular, human observers were not only more robust to this manipulation (Figure 2), but also their misclassifications were not well matched by any of the models (Figure 3, bottom row). The ratios for the two color manipulations (not plotted) were less than +/-2% from unity across all classifiers, and the results for the block-scramble-with-grid manipulation were again similar to the basic scramble condition (Figure S4). To examine the agreement levels in more detail, we next selected a subset of image degradations to analyze classification confidence on an image-by-image basis.

In Figure 6, we plot measures of "confidence" for the InOutNet, GIST-LDA, and HSV-LDA classifiers, versus the probability of the human observer responses (i.e., the confidence of the crowd) for several levels of the blur manipulation. Data for PlacesNet are not shown, because top-five classification was used for this classifier, rather than the softmax output. Each panel contains the data for a different level of blur (starting from un-blurred and increasing to the right). Data points are plotted separately for outdoor (light circles) and indoor (dark triangles) images. Black X's on each axis indicate the average response, providing a measure of overall bias. In these plots, points falling along the diagonal would indicate similar classification and confidence between the human observers (abscissa) and the classifier (ordinate). Data within the gray regions indicate similar classification, but not necessarily similar confidence. For the InOutNet (upper row), increasing levels of blur resulted in an increasing number of indoor images (triangles) being classified as outdoors. This is also true for the GIST-LDA classifier (middle row). At the highest levels of blur, both classifiers categorized all images as outdoors, with high confidence. This pattern was not mirrored by the human observers, who remained relatively unbiased across all blur levels. We might hypothesize that the lack of high spatial frequencies resulting from blurring is causing this

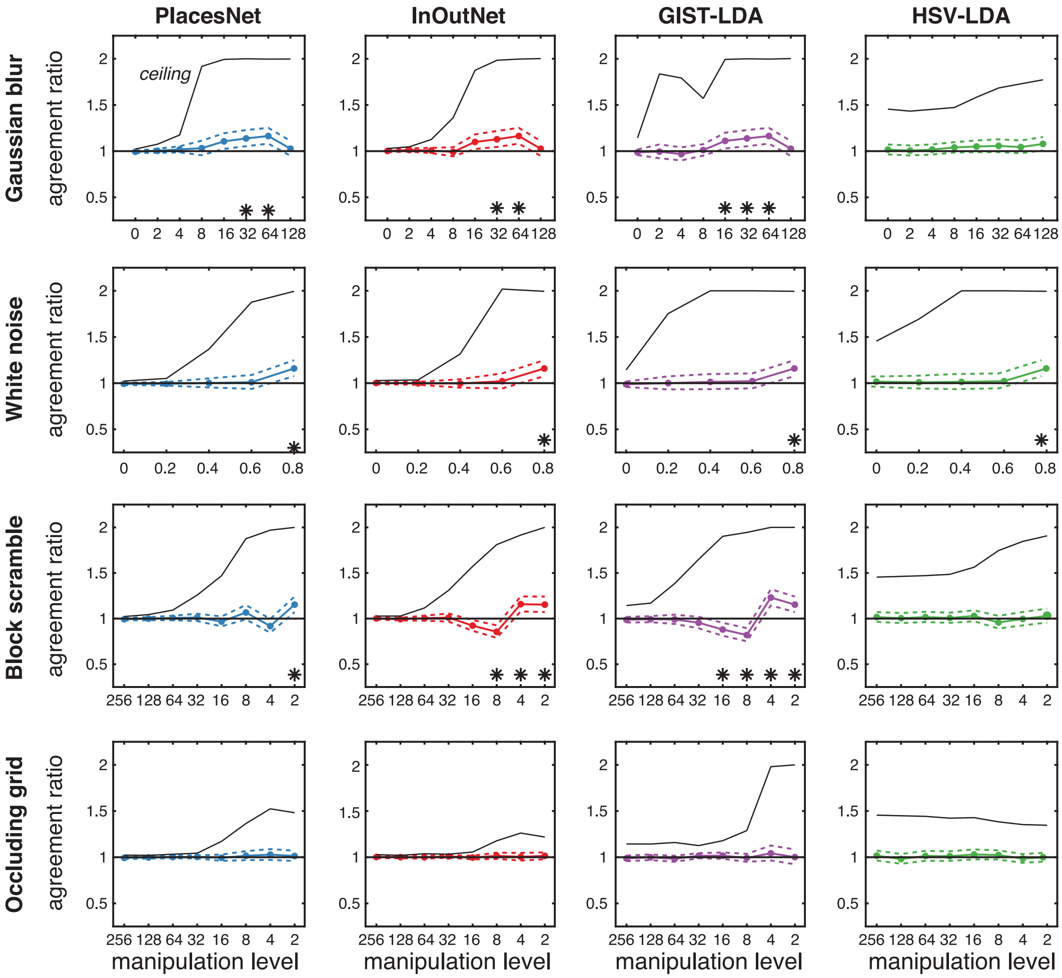

Fig. 5. Agreement ratio of each classifier with human observers, separated by manipulation type and level. Data are plotted in the same manner as Figure 4, except that expected agreement was calculated within each manipulation and level only. Ratios significantly different from one are indicated with asterisks. Units for manipulation levels (abscissa) are the same as in Figure 2.

breakdown. However, a similar pattern occurred at the highest levels of scrambling and noise (Figures S5–7 in the Supplementary Material). Because the training dataset was balanced in the number of indoor and outdoor examples, it is unclear why extremely scrambled/noisy and extremely blurred images would result in outdoor classifications with high confidence. However, we hypothesize that the absence of mid-range spatial scales may provide a strong cue that an image is outdoors. This cue could be a direct consequence of the fact that outdoor environments tend to have a larger spatial extent than indoor environments, and that differences in spatial scale have known differences in amplitude spectra (Torralba and Oliva 2003). For comparison to the previous agreement analysis, above each panel in Figure 6, we indicate the corresponding agreement ratio from Figure 6 for that subset of images. Of the conditions plotted, the only one with a significant agreement ratio is the 16-pixel blur kernel with the GIST-LDA classifier. In this case, the human observers and classifier appear to have a bias to classify images as outdoors. Overall, the patterns observed in this confidence analysis support the results shown

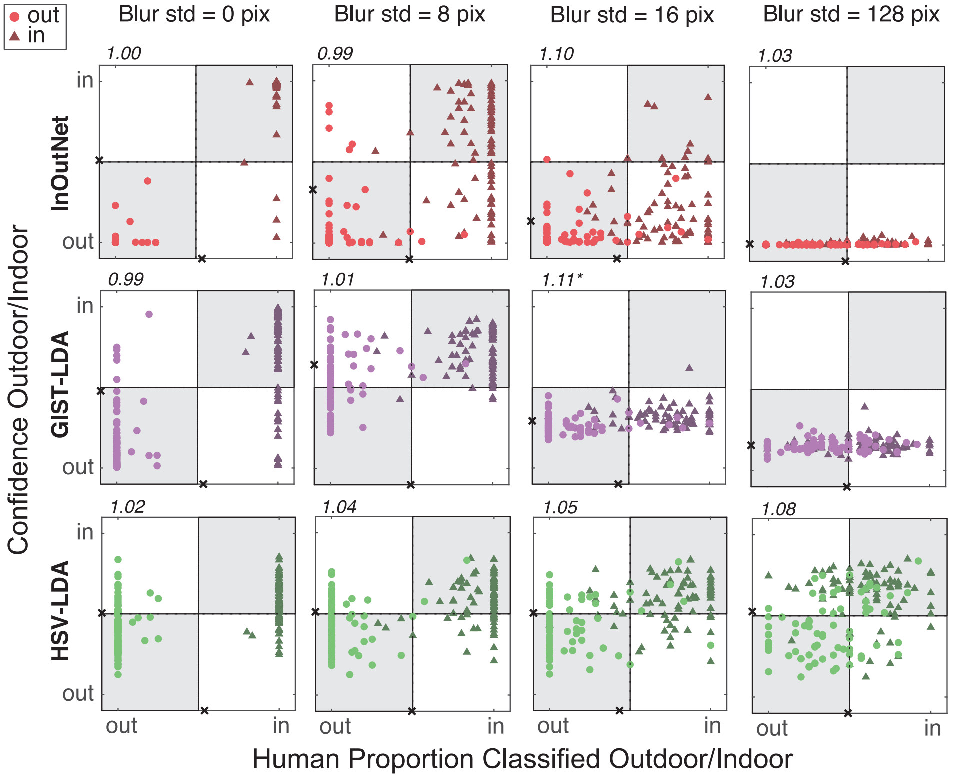

Fig. 6. Class probabilities and human classifications for the blur condition. Each panel shows the data for a single Gaussian blur kernel standard deviation (columns) and classifier (rows). Data points in each panel show the proportion of indoor/outdoor responses from the human observers (abscissa) and classifier confidence (ordinate) for an individual test image, ranging from −1.0 outdoor to 1.0 indoor. Data points are plotted separately for outdoor (bright circles) and indoor (dark triangles) images. Black X's on each axis indicate the average response across each subset of images. Italicized values above each panel correspond to the agreement ratio associated with these data, with statistically significant values followed by an asterisk. In = indoor; out = outdoor; std = standard deviation.

in the criterion analysis (Figure S2), which also indicated that the classifiers tended to have strong biases for degraded images that were not reflected in the human responses.

In summary, our analysis suggests decreases in classifier accuracy associated with image degradations tended to also be associated with increased biases. Similar to previous work on object classification, the biases were at times extreme and not shared with human observers (Dodge and Karam 2019). This result suggests that the classifiers' representations of scene properties may share similar features to human perceptual representations, but for extremely degraded images this comparison seems to break down, because the classifiers have an overriding bias not reflected in human observers. A clear exception to this is the HSV-LDA classifier (Figure 6, bottom row), which has relatively consistent performance and bias across many manipulation levels. However, this classifier is not very accurate.

In a final exploratory analysis, we examined the level of agreement between the two CNN-based classifiers. The results of this analysis are shown in Figure 7. This is an agreement analysis between PlacesNet and InOutNet, plotted in the same manner as Figure 5. The two networks agreed strongly with each other on their classifications of the blurred and noisy images, but surprisingly diverge in their classifications of the mid-level scrambled

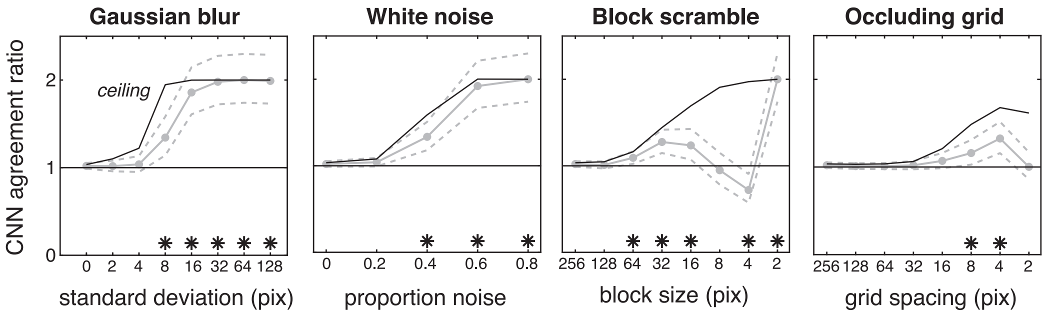

Fig. 7. Agreement ratio between PlacesNet and InOutNet, separated by manipulation type and level. Data are plotted in the same manner as Figure 5. Note that in some cases the 95% confidence interval on the ratios exceeds our plotted ceiling. This is because the confidence interval is calculated generally for the observed ratio of proportions (Equation 3), whereas the ceiling shows the maximum possible agreement ratio assuming that the expected value is fixed (Equation 6). Ratios significantly different from one are indicated with asterisks.

images. This result suggests that the local features learned by the PlacesNet network trained to classify subcategories are substantially different from those learned by InOutNet. For the grid condition, there appears to be some inter–classifier agreement.

5 DISCUSSION

Deep CNNs have emerged as a computer vision tool both for practical applications and for modeling biological visual processing. For both of these scenarios, it is important to quantify and understand classification performance from non-ideal images. For computer vision applications, this question is important for characterizing how such classifiers may work "in the wild." For understanding human vision, analyzing the pattern of errors can be a useful way for examining the extent of representational similarity between biological and computer vision systems. The current study addressed these issues for scene recognition, a visual task that is ubiquitous both for computer and biological vision.

Whenever computer vision performance is quantified, it is essential to understand how performance compares to established upper and lower bounds. In the current project, we defined a set of human observers on Amazon Mechanical Turk as our presumed upper bound, and a simple classifier based on global image hue, saturation, and value as the lower bound. On intact scene images, none of our classifiers out-performed human observers. On a practical level, however, accuracy of the CNNs remained well above chance for a wide range of image manipulations that disrupted both local and global image statistics. Surprisingly, the only conditions in which humans were surpassed were the high blur and scramble conditions, and this was obtained with the presumed lower-bound HSV-LDA classifier. This result suggests that a computer vision approach to scene classification that combines hierarchical feature processing (CNNs) and weakly informative global scene cues may provide a powerful and robust tool for automated scene classification.

Towards the second aim of understanding human scene category representations by examining the similarities and differences with classifiers, we can make several observations: The first is that human observers were not similarly robust to all types of image manipulations: high levels of blur and scramble were more difficult for observers than high levels of white noise or a narrow occluding grid. The latter two manipulations maintain the extended scene contours and relational information across local regions that have been shown to contribute to scene classification ability (Biederman 1972; Walther and Shen 2014). Strikingly, we observed that none of the four classifiers tested here showed any substantial agreement with human observers on degraded images, suggesting that humans and classifiers are not using the available visual information in the same way. As the grid manipulation replaces pixels with the average RGB value, this manipulation can be thought of as introducing new high-frequency horizontal and vertical information to the image. GIST features come from outputs of oriented filters, so this new information is disruptive to classification. Interestingly, the CNN-based classifiers were relatively unimpaired until the imposed grid became as fine as 32 pixels (PlacesNet), or 16 pixels (InOutNet), corresponding to receptive field sizes that are intermediate to the first and second convolutional layers of these networks. As early human vision is also believed to use similar filtering operations, we are faced with the need to find the subsequent operations in biological vision that allow us to transcend noise from these early representations. As CNNs are exclusively feedforward models, feedback from higher cortical areas could be implicated. By contrast, the two CNNs showed more agreement with human observers when faced with blurred images, suggesting that both humans and CNNs may utilize high spatial frequency information similarly when performing scene classification.

It is important to consider how the current results may generalize to other high-level visual tasks involving scenes, objects, and people. Previous work using a wide variety of scene-recognition approaches showed that classification of outdoor categories (e.g., coast, forest, highway) using scrambled images was on average better and more correlated to human responses than classification of indoor categories (e.g., bathroom, bedroom, kitchen) (Borji and Itti 2014). It has been suggested that indoor scenes are fundamentally more variable in their local visual appearance (Parikh 2011), and may thus require human observers to call upon higher-level visual processing that is not well modeled with existing computer vision approaches. Indeed, our own results comparing PlacesNet and InOutNet suggest that the two networks diverge in their representation of local image properties. With respect to other tasks, recent work on object classification suggests neural networks trained for object classification are relatively un-representative of both human categorizations and neural activity measured in macaque visual cortex when examined on an image-by-image basis (Rajalingham et al. 2018). There is thus converging evidence that the current generation of CNNs have high agreement with humans on average (Greene et al. 2016; Jozwik et al. 2017; Kubilius et al. 2016), but poor agreement on the scale of individual images. However, more work is needed to examine this possibility across a range of visual tasks.

Although we have shown that the classifiers' response patterns tend to diverge from humans for highly degraded images, it is possible that this pattern may be altered if the classifiers are explicitly trained to handle degraded images. Prior work has shown that both fine-tuning and re-training CNNs with blurry and noisy images increases the classification accuracy on these image degradations (Dodge and Karam 2019; Zhou et al. 2017). In fact, Vasiljevic et al. (2016) showed that CNNs fine-tuned with one type of blur can generalize to accurately classify other types of blurred images. These results suggest that CNNs are capable of classifying degraded images accurately and generalizing to other types of (similar) degradations. Another recent study compared CNNs and humans, and found that CNNs can outperform humans when performing object recognition from degraded images when the degradation is included in training (Geirhos et al. 2018). This study contained a wide variety of image manipulations, and also found that training can generalize across degradation types in some cases, but in other cases it clearly does not. Human performance was overall more robust across a wide range of image manipulations. This result suggests that while training on degraded images can increase CNN accuracy, it may not increase representational similarity to the human perceptual system. Future work may examine whether there is some combination of different image manipulations that can be used in training to collectively produce more human-like visual representations.

It is particularly interesting to consider these issues of accuracy and agreement when it comes to cases in which both human and computer vision may be combined to achieve a common goal. For example, artificial-intelligence-based vision aids for people who are blind or visually impaired have been proposed (and some are available) that use computer vision to assist with interpreting information in natural scenes (e.g., (Everingham et al. 2003; Manduchi and Coughlan 2012; Manduchi et al. 2010; Microsoft n.d.; Wang et al. 2014; Zhao et al. 2016)). In particular, a camera on a mobile device or head-mounted display may be used to automatically parse information related to the surrounding scene that is difficult for the user to detect. This information could then be conveyed via automated speech or other modalities. The current study suggests that CNN-based scene classifiers can perform reasonably well on blurry and noisy images, which may be encountered from mobile device cameras. In addition, the low agreement with human classifications suggests that information from a CNN might provide complementary information that can be combined with other cues that humans rely on. However, the manner in which scene images are degraded in this real-world use case also likely differs from the specific manipulations tested in the current work. Motion blur, for example, may be more common than Gaussian blur. Moreover, it is important for classification performance to generalize not just to noisy images, but also to non-canonical viewpoints that may not be well represented in current image sets (e.g., scenes falling into multiple categories, camera angles oriented in different directions).

For real-world applications that rely on both human and computer vision, it will be important in the future to thoroughly examine how a classifier trained on images from the internet performs in real-world use cases, and to explore how scene information can be conveyed in a way that maximizes the usefulness for the user. Extending to more dynamic, real-world training sets may further improve performance on degraded images, and at the same time may serve to narrow the gap between human and computer visual scene representations.

6 SUPPLEMENTARY MATERIALS

See the supplementary materials in the online version. In addition, raw classification data, the InOutNet classifier, and analysis scripts in Matlab associated with this manuscript are publicly available at https://osf.io/q4xtc/.

REFERENCES

Pulkit Agrawal, Dustin Stansbury, Jitendra Malik, and Jack L. Gallant. 2014. Pixels to voxels: Modeling visual representation in the human brain. Retrieved from: *Arxiv Preprint Arxiv:1407.5104* (2014), 15.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Royal Stat. Soc. Series B (Methodol.)* 57, 1 (1995), 289–300.

Irving Biederman. 1972. Perceiving real-world scenes. *Science* 177, 4043 (1972), 77–80.

Irving Biederman, Arnold L. Glass, and E. Webb Stacy. 1973. Searching for objects in real-world scenes.*J. Exper. Psychol.* 97, 1 (1973), 22–27.

Ali Borji and Laurent Itti. 2014. Human vs. computer in scene and object recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 113–120.

Muriel Boucart, Christine Moroni, Miguel Thibaut, Sebastien Szaffarczyk, and Michelle Greene. 2013. Scene categorization at large visual eccentricities. *Vis. Res.* 86 (2013), 35–42.

Charles F. Cadieu, Ha Hong, Daniel L. K. Yamins, Nicolas Pinto, Diego Ardila, Ethan A. Solomon, Najib J. Majaj, and James J. DiCarlo. 2014. Deep neural networks rival the representation of primate IT cortex for core visual object recognition. *PLoS Comput. Biol.* 10, 12 (2014), e1003963.

Radoslaw M. Cichy, Aditya Khosla, Dimitrios Pantazis, Antonio Torralba, and Aude Oliva. 2016. Deep neural networks predict hierarchical spatio-temporal cortical dynamics of human visual object recognition. Retrieved from: *Arxiv Preprint Arxiv:1601.02970* (2016), 15.

Samuel Dodge and Lina Karam. 2016. Understanding how image quality affects deep neural networks. In *Proceedings of the 8th International Conference on Quality of Multimedia Experience (QoMEX'16).* IEEE, 1–6.

Samuel Dodge and Lina Karam. 2019. Human and DNN classification performance on images with quality distortions: A comparative study. *ACM Trans. Appl. Percept.* 16, 2, Article 7 (Mar. 2019), 17 pages.

M. R. Everingham, Barry T. Thomas, and Tom Troscianko. 2003. Wearable mobility aid for low vision using scene classification in a Markov random field model framework. *Int. J. Human-Comput. Interact.* 15, 2 (2003), 231–244.

Kunihiko Fukushima. 1988. Neocognitron: A hierarchical neural network capable of visual pattern recognition. *Neural Netw.* 1, 2 (1988), 119–130.

Robert Geirhos, Carlos R. M. Temme, Jonas Rauber, Heiko H. Schütt, Matthias Bethge, and Felix A. Wichmann. 2018. Generalisation in humans and deep neural networks. In *Adv. Neural Inform. Proc. Syst.* 7549–7561.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. Retrieved from: *Arxiv Preprint Arxiv:1412.6572* (2015), 1–11.

Michelle R. Greene. 2013. Statistics of high-level scene context. *Front. Psychol.* 4 (2013), 777.

Michelle R. Greene, Christopher Baldassano, Andre Esteva, Diane M. Beck, and Li Fei-Fei. 2016. Visual scenes are categorized by function.*J. Exper. Psychol.: Gen.* 145, 1 (2016), 82–94.

Michelle R. Greene and Bruce C. Hansen. 2018. Shared spatiotemporal category representations in biological and artificial deep neural networks. *PLoS Comput. Biol.* 14, 7 (2018), e1006327.

Michelle R. Greene and Aude Oliva. 2009a. The briefest of glances: The time course of natural scene understanding. *Psychol. Sci.* 20, 4 (2009), 464–472.

Michelle R. Greene and Aude Oliva. 2009b. Recognition of natural scenes from global properties: Seeing the forest without representing the trees. *Cog. Psychol.* 58, 2 (2009), 137–176.

Iris I. A. Groen, Michelle R. Greene, Christopher Baldassano, Li Fei-Fei, Diane M. Beck, and Chris I. Baker. 2018. Distinct contributions of functional and deep neural network features to representational similarity of scenes in human brain and behavior. *Elife* 7 (2018), e32962.

Iris I. A. Groen, Edward H. Silson, and Chris I. Baker. 2017. Contributions of low- and high-level properties to neural processing of visual scenes in the human brain. *Phil. Trans. R. Soc. B* 372, 1714 (2017), 20160102.

Umut Güçlü and Marcel A. J. van Gerven. 2015. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *J. Neurosci.* 35, 27 (2015), 10005–10014.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70.* JMLR.org, 1321–1330.

Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. 2014. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM International Conference on Multimedia.* ACM, 675–678.

Kamila M. Jozwik, Nikolaus Kriegeskorte, Katherine R. Storrs, and Marieke Mur. 2017. Deep convolutional neural networks outperform feature-based but not categorical models in explaining object similarity judgments. *Front. Psychol.* 8 (2017), 1726.

Samil Karahan, Merve K. Yildirum, Kadir Kirtac, Ferhat S. Rende, Gultekin Butun, and Hazim K. Ekenel. 2016. How image degradations affect deep CNN-based face recognition? In *Proceedings of the International Conference of the Biometrics Special Interest Group (BIOSIG'16).* 1–5.

D. Katz, J. Baptista, S. P. Azen, and M. C. Pike. 1978. Obtaining confidence intervals for the risk ratio in cohort studies. *Biometrics* 34, 3 (1978), 469–474.

Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. 2014. Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Comput. Biol.* 10, 11 (2014), e1003915.

P. A. R. Koopman. 1984. Confidence intervals for the ratio of two binomial proportions. *Biometrics* 40, 2 (1984), 513–517.

Nikolaus Kriegeskorte. 2015. Deep neural networks: A new framework for modeling biological vision and brain information processing. *Ann. Rev. Vis. Sci.* 1 (2015), 417–446.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A. Bandettini. 2008. Representational similarity analysis-connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* 2 (2008), 4.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Proceedings of the Conference on Advances in Neural Information Processing Systems 25.* 1097–1105.

Jonas Kubilius, Stefania Bracci, and Hans P. Op de Beeck. 2016. Deep neural networks as a computational model for human shape sensitivity. *PLoS Comput. Biol.* 12, 4 (2016), e1004896.

Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. 2006. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06).* IEEE, 2169–2178.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature* 521, 7553 (2015), 436–444.

Li-Jia Li, Hao Su, Li Fei-Fei, and Eric P. Xing. 2010. Object bank: A high-level image representation for scene classification & semantic feature sparsification. In *Proceedings of the Conference on Advances in Neural Information Processing Systems 23.* 1378–1386.

Roberto Manduchi and James Coughlan. 2012. (Computer) vision without sight. *Commun. ACM* 55, 1 (2012), 96–104.

Roberto Manduchi, Sri Kurniawan, and Homayoun Bagherinia. 2010. Blind guidance using mobile computer vision: A usability study. In *Proceedings of the 12th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS'10).* ACM, New York, NY, 241–242.

David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. 2001. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings of the 8th IEEE International Conference on Computer Vision (ICCV'01)*, Vol. 2. IEEE, 416–423.

Microsoft. [n. d.]. Seeing AI. Retrieved from: https://www.microsoft.com/en-us/seeing-ai.

Aude Oliva and Antonio Torralba. 2001. Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. J. Comput. Vis.* 42, 3 (2001), 145–175.

Devi Parikh. 2011. Recognizing jumbled images: The role of local and global information in image classification. In *Proceedings of the IEEE International Conference on Computer Vision.* 519–526.

Matthew F. Peterson, Jing Lin, Ian Zaun, and Nancy Kanwisher. 2016. Individual differences in face-looking behavior generalize from the lab to the world. *J. Vis.* 16, 7 (2016), 12–12.

Rishi Rajalingham, Elias B. Issa, Pouya Bashivan, Kohitij Kar, Kailyn Schmidt, and James J. DiCarlo. 2018. Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *J. Neurosci.* 38, 33 (2018), 7255–7269.

Laura Walker Renninger and Jitendra Malik. 2004. When is scene identification just texture recognition? *Vis. Res.* 44, 19 (2004), 2301–2311.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein et al. 2015. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* 115, 3 (2015), 211–252.

Harold Stanislaw and Natasha Todorov. 1999. Calculation of signal detection theory measures. *Behav. Res. Meth., Instr., Comput.* 31, 1 (1999), 137–149.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. Retrieved from: *Arxiv Preprint Arxiv:1312.6199* (2013), 1–10.

Benjamin W. Tatler. 2007. The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *J. Vis.* 7, 14 (2007), 1–17.

Antonio Torralba. 2009. How many pixels make an image? *Vis. Neurosci.* 26, 1 (2009), 123–131.

Antonio Torralba and Alexei A. Efros. 2011. Unbiased look at dataset bias. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'11)*. IEEE, 1521–1528.

Antonio Torralba and Aude Oliva. 2003. Statistics of natural image categories. *Netw.: Comput. Neural Syst.* 14, 3 (2003), 391–412.

Barbara Tversky and Kathleen Hemenway. 1983. Categories of environmental scenes. *Cog. Psychol.* 15, 1 (1983), 121–149.

Igor Vasiljevic, Ayan Chakrabarti, and Gregory Shakhnarovich. 2016. Examining the impact of blur on recognition by convolutional networks. Retrieved from: *Arxiv Preprint Arxiv:1611.05760* (2016), 10.

Julia Vogel, Adrian Schwaninger, Christian Wallraven, and Heinrich H. Bülthoff. 2007. Categorization of natural scenes: Local versus global information and the role of color. *ACM Trans. Appl. Percept.* 4, 3 (2007), 19–es.

Dirk B. Walther and Dandan Shen. 2014. Nonaccidental properties underlie human categorization of complex natural scenes. *Psychol. Sci.* 25, 4 (2014), 851–860.

Shuihua Wang, Hangrong Pan, Chenyang Zhang, and Yingli Tian. 2014. RGB-D Image-based detection of stairs, pedestrian crosswalks and traffic signs. *J. Vis. Commun. Image Rep.* 25, 2 (2014), 263–272.

David M. Watson, Timothy J. Andrews, and Tom Hartley. 2017. A data driven approach to understanding the organization of high-level visual cortex. *Sci. Rep.* 7, 1 (2017), 3596.

Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. 2010. Sun database: Large-scale scene recognition from abbey to zoo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'10)*. IEEE, 3485–3492.

Daniel L. K. Yamins and James J. DiCarlo. 2016. Using goal-driven deep learning models to understand sensory cortex. *Nature Neurosci.* 19, 3 (2016), 356–365.

Daniel L. Yamins, Ha Hong, Charles Cadieu, and James J. DiCarlo. 2013. Hierarchical modular optimization of convolutional networks achieves representations similar to macaque IT and human ventral stream. In *Proceedings of the Conference on Advances in Neural Information Processing Systems 26*. 3093–3101.

Yuhang Zhao, Sarit Szpiro, Jonathan Knighten, and Shiri Azenkot. 2016. CueSee: Exploring visual cues for people with low vision to facilitate a visual search task. In *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp'16)*. ACM, New York, NY, 73–84.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2921–2929.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2018. Places: A 10 million image database for scene recognition. *IEEE Trans. Pattern Anal. Machine Intell.* 40, 6 (2018), 1452–1464.

Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. 2014. Learning deep features for scene recognition using places database. In *Proceedings of the Conference on Advances in Neural Information Processing Systems 27*. 487–495.

Yiren Zhou, Sibo Song, and Ngai-Man Cheung. 2017. On classification of distorted images with deep convolutional neural networks. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'17)*. IEEE, 1213–1217.

Received October 2018; revised April 2019; accepted May 2019