5-30-2017

Scene classification from degraded images: comparing human and computer vision performance

Tim M. Tadros
Dartmouth College

Dartmouth College Computer Science Technical Report TR2017-820

Scene classification from degraded images: comparing human and computer vision performance

Tim Tadros
Advised by Emily Cooper

May 30, 2017

Abstract

People can recognize the context of a scene with just a brief glance. Visual information such as color, objects and their properties, and texture are all important in correctly determining the type of scene (e.g. indoors versus outdoors). Although these properties are all useful, it is unclear which features of an image play a more important role in the task of scene recognition. To this aim, we compare and contrast a state-of-the-art neural network and GIST model with human performance on the task of classifying images as indoors or outdoors. We analyze the impact of image manipulations, such as blurring and scrambling, on computational models of scene recognition and human perception. We then create and analyze a measure of local-global information to represent how each perceptual system relies on local and global image features. Finally, we train a variety of neural networks on degraded images to attempt to build a neural network that emulates human performance on both classificaton accuracies and this local-global measure.

1 Introduction

Scene recognition is a difficult and important task in both biological and computer vision. However, in the domain of computer vision research, comparatively little emphasis has been placed on scene recognition, despite its potential utility for providing context for many visual tasks, including object recognition [38]. Objects rarely occur in isolation and the general context in which they appear may be helpful in recognizing certain objects. Indeed, in people, scene recognition probably occurs in parallel with object detection and both tasks influence each other [10]. Moreover, scene recognition may be important in enabling people to interpret their environment, which affects subsequent behavior and how stimuli such as the objects in the scene are perceived [6] [3]. For this reason, we look at scene recognition in human and machine vision to study how both people and computational systems perform this complex task.

If image information is systematically degraded, can computer vision models infer the lost visual information as well as people can? Studying which features are important for human beings may help us improve the robustness and invariances of deep convolutional neural networks trained to perform scene recognition.

To remove or enhance either local or global image properties, images can be manipulated to contain more of one feature than another. To eliminate global information, we can segment an image into various blocks of different sizes and scramble those blocks to different regions of the image. It has been shown that jumbling an image decreases object recognition, likely by limiting contextual information available to the perceiver [3]. The spatial layout of the image, such as its navigabiliity and openness, is lost but, depending on the size of the blocks, much of the local information about objects is preserved. Likewise, we can remove local information but preserve global information by adding noise or blurring an image. With these techniques, objects become harder to discern but the spatial layout of the scene remains intact. In order to determine if people use more local or global information when they perform this scene recognition task, we can measure their performance on classifying the type of scene depicted in degraded images. This will create a causal link between the perception of the scene and the image features available.

2 Human Scene Recognition

The human visual pathway has a hierarchichal structure, where low-level features are computed first and higher-level conceptual details are computed later on. First, low-level vision is responsible for the extraction of low-level features, such as depth, color, and texture as well as representing certain surfaces and edges [17]. Then, higher-level vision is responsible for mapping these low-level features to meaning, such as recognition of conceptual scenes and objects [12]. Given this hierarchical structure, the question of what features scene recognition relies on arises. It may be the case that scene recognition depends on the detection of objects in those scenes or, on the contrary, it may be that people can recognize scenes without first determining which objects are in the scene. Scene identification research has mainly focused on the time course of interdependent visual tasks and the features in scenes that play a role in scene recognition.

Human scene recognition was first studied mainly by looking at how quickly people can recognize a scene. In one study, researchers asked people to identify a target scene in a slideshow of images, where each image was presented for 113 ms [25] [26]. When a verbal description of the target scene was displayed before the sequence of images was presented, people were as accurate at selecting the target image as when they had seen the target image beforehand. This suggests that the context of a scene can be understood within 113 ms. At 500 ms, people are nearly perfect at recognizing if a given scene is indoors or outdoors. At shorter presentation times, however, people were more likely to categorize scenes as outdoors than indoors, suggesting a slight bias for outdoor scenes when stimuli are presented only briefly [8].

Other human scene recognition research has looked at the role of objects in performing scene recognition. It was commonly thought that objects in an image played a role in identifying the overall context, or scene class, of an image. In one study, it was proposed that objects are responsible for activating a schema of the scene which would then lead to more advanced scene recognition [2] (see [12] and [10] for a review of this view). This view treats objects as the building blocks of recognition tasks and suggests that objects are recognized before scenes.

An opposing view, the scene-centered view, treats the scene as a whole as the building block for complex recognition tasks. Rather than recognizing the objects first, people get an overall sense of the scene and then either use the context of the scene to identify the objects within the scene, or perform scene and object recognition in parallel. In conditions where objects are not easily identifiable, such as low frequency images or images with sparse contours, people can still accurately perform scene recognition [21] [13]. This suggests that scene identity may be computed before or in parallel with object identity.

2.1 Reliance on Local and Global Features

Scene recognition in people has also been looked at in the context of scene information, ranging from local to global information. In the literature, it is assumed that both the local and global properties of a scene are important in enabling people to perform contextual scene recognition [28] [19]. An image's local properties include the objects in the image as well as their features, such as line orientation, color, and texture. Support for local processing comes from the way the human visual system is setup to perform visual processing. One study looked at EEG data of people observing visual scenes to find that the earliest components in EEG signals that are related to recognition have an onset of 150-300 ms. There are other EEG signals that are measured after 130 ms from the onset of the visual stimulus and these signals correlate to low-level visual features in the image [14]. Since low-level features are processed before higher-level features that correspond to the layout of the scene, and scene recognition itself is a rapid process, it seems that low-level features could be more indicative of scene category. Although, low-level features are processed first, it is not clear from this study how these low-level features influence scene recognition, since there are other factors that seem to influence visual recognition more directly (those that come after 150 ms).

Another study looked at the role of texture in scene recognition [28]. Researchers conducted a perceptual study and built a computational model tuned to textural features in images to determine if their models could be a good measure of people's performance on the perceptual study. They found that, when subjects are presented with images for only a brief period of time (less than 100 ms), the texture model mimics human performance. This suggests that low-level features such as texture could be important in understanding the gist of a scene. Another study on local image features identified the role of color as being important when it is diagnostic of scene category [20]. Researchers found that when color is not relevant in identifying an image's scene category, then subjects do not have a delayed response in identifying abnormally colored images. However, when color is pertinent (such as in a desert scene), then subjects were much slower to identify images presented as abnormally colored. These studies suggest that local information, such as object color and texture, do play an important role in scene recognition.

Although local information seems important in scene classification, when local information in one patch of an image contradicts local information in another patch, people have difficulties performing scene recognition. In one study, researchers observed that the spatial configuration of objects is important in detecting objects. If two objects are present in isolation, then subjects have no trouble identifying those objects. However, when the objects are put together and improperly spaced (based on a subject's expectations), then it takes much longer to recognize those objects. Likewise, when ambiguous objects are placed with clearly recognized objects, then the ambiguous objects are more quickly and accurately identified than when they are placed next to other ambiguous objects [1]. These results suggest that the global representation of objects in a scene may also be important in identifying the context of a scene, or other objects within the scene.

It seems likely that while local properties may be important in quick scene categorization, more accurate categorization is done through a combination of local and global properties. An image's global properties include the openness, navigability, concealment of objects, depth, and expansion within the image [10] as well as the spatial layout and construction of the objects within the scene [23]. Openness represents the spatial enclosement of the scene. A crowded room, for example, would be relatively less open than a field since there are many objects obstructing the field of view of a bystander looking at the scene. Navigability corresponds to the ease at which one could move through the scene. High levels of concealment indicate that many objects are hidden within the scene. The depth of a scene refers to the size of the space; a single room would have a small depth while a mountain range would have a very large depth. In sum, these properties refer to how objects in a scene are held together and how one may interact with the scene. They do not measure what types of objects are in the scene and what properties those objects possess.

Global properties have been shown to be very useful in scene categorization. In one study, researchers had participants rank images along seven scales that measured global properties (these scales were openness, expansion, mean depth, navigability, temperature, transience, and concealment). After ranking these images, a different set of participants was asked to map each image to a scene category. It was hypothesized that images from different scenes that were close together on the seven dimensions of global properties would be confused for each other more often than images that were farther apart [10]. Researchers confirmed this hypothesis, which suggests that global image information is important, and may be primary, in scene recognition.

Another study addressed the importance of the background of an image in performing object recognition. When an object is presented with a coherent background (so the context is expected), object detection is easier. However, when that same object is presented with an incoherent background, performance on the object recognition task dropped by around ten percent [5]. Although they studied the effects of global information on object detection, it seems that understanding the context of the scene aids in object detection. This means that scene information may be captured before object information.

Other studies found that people can learn the global features of a scene without being aware of the local features. However, the opposite, being aware of objects without understanding the context of those objects is more difficult [18]. Whether human vision utilizes more local or global visual processing is an important question. If people rely on global processing for scene categorizaiton, then this suggests that people get an overall sense of the scene before they direct their attention to individual objects in the scene. If local information is more important in scene categorization, then it seems that both object and scene information are captured in parallel.

3 Scene Recognition in Machine Vision

3.1 Reliance on Local/Global Features

Determining which features are important for the human visual system to perform scene recognition may help us improve the performance of scene recognition computational models. There have been many attempts at building computational scene recognition systems in both classical machine learning and deep learning. Some of these classical approaches calculate a combination of local and global information and use this to ground their higher level scene categorization.

The GIST model [22] is a classical machine learning approach to scene recognition. It converts each image in the training set into a pre-determined feature vector and uses this feature vector to train a classifier. The GIST model computes a feature vector based on low-level features in each image, termed the “spatial envelope”, which represents the “naturalness, openness, roughness, expansion, ruggedness” of the scene, similar to the seven global dimensions explained above. These dimensions are found to describe the variation in most scene images and thus have been shown to be successful in building a classifier trained for scene recognition. For example, natural scenes can be distinguished from man-made scenes based on the absence of sharp horizontal and vertical lines and the presence of grainy texture. Likewise, scenes with a horizon line are usually very open, while scenes without horizon lines and with many smaller lines throughout are usually enclosed. Rugged scenes usually obscure the horizon line and contain many contours. Computing these features can help map an image along dimensions of ruggedness, openness, naturalness, and expansion. Then, a classifier can learn how these features relate to scene classes and use these dimensions to predict an image's class. The GIST model successfully classified 82% of indoor and outdoor scenes, which suggests that these global features are indicative of an image's indoor or outdoor class.

Another model looked at the role of local information in scene classification [35]. Researchers investigated the individual object classes that compose each scene type. For example, the scene class "forest" can be thought of as composed of various objects, such as the sky, water, grass, trunks, foliage, etc. Each scene class contains a different percentage of each object. By computing prototypical representations of each scene class and which object components are especially discriminant for a scene class, an image can be mapped to a scene class based on its proximity to different prototypical representations. By looking at the percentage of the image composed of each type of object, this model looks at local, object-based information rather than the spatial relationships of those objects. It was found that this model has a classification accuracy of 90% when objects were annotated before classification and 67% when objects were annotated using machine learning.

Other local property-based scene classification methods include looking at line orientation to differentiate between city and landscape scenes. City scenes contain many vertical lines representing buildings while landscape scenes lack such vertical orientation [9]. Analyzing the orientation of similar textures in a scene can lead to classification properties that distinguish two scene categories. Other local-based models look at the role of texture, color, and frequency to perform indoor-outdoor scene classifcation [31]. A model that computed these features in subcomponents of an image and then stacked these features performed the best with a classification accuracy of 90% on the indoor-outdoor classification task.

Both local- and global-based computational models perform similarly on the indoor-outdoor classification task. It has also been noted that many existing computational models are either exclusively trained at classifying outdoor scene classes or biased toward outdoor scene classes [27]. This may be because there has not been a large indoor scene class database. Despite machine vision systems' strong performance on the indoor-outdoor classification task, they are not yet as accurate as people.

3.2 Deep Convolutional Neural Networks

Neural networks improve upon these single layer classification models by adding several non-linearities to compute new features before performing classification. For images, convolutional neural networks utilize the layout of the image and compute different features for each patch of an image in order to create a part-based representation of the image. Each unit in a convolutional neural network responds to pixels in a small region of the input image, known as the receptive field. It has been shown that each unit in a convolutional neural network can be trained to respond to different features in the receptive field, depending on the values of the parameters in the filter associated with that unit. This information can be used by neurons deeper in the network to determine if the image has certain shapes or objects. By building up multiple layers, neural networks create a more complex representation of the image that can be useful for scene recognition.

Deep convolutional neural networks have reached a performance level on many visual tasks that is comparable to or exceeds human performance. Convolutional neural networks perform comparably to humans on tasks such as facial recognition and handwritten digit recognition [32] [30] [37]. They have also passed human performance on some object recognition tasks [11]. However, it does not seem that neural networks are as robust as people when it comes to interpreting new images. In one study, a state-of-the-art face detection system was fooled by adding eyeglass frames to the faces. The neural network was unable to correctly identify these faces when there was an external accessory in the image [29].

A state-of-the-art scene neural network trained on scene recognition currently achieves accuracies of around 95% on intact images on the indoor/outdoor classification task [40]. This network, AlexNet trained on the Places database, is trained on 205 scene classes and achieves a classification accuracy of 56.2% on this database. Since neural networks have surpassed many of the traditional machine learning solutions to vision tasks and have been designed with the human brain in mind, it seems promising to look at convolutional neural networks when examining any visual task.

4 Comparisons Between Human and Computer Vision

Three previous studies have looked at human and computer performance on scene recognition of jumbled images [4], [36] [24]. In one study, the authors found that human performance on recognizing jumbled images falls as the blocksize of the jumbled patches in the image decreases [4]. Using 14 computational models built for both scene and object classification, they measured various models' performances on classifying images scrambled into 6x6 pixel blocksizes. All computational models are trained as one-vs-all support vector machines utilizing different feature vectors. They used multiple scene categories for indoor and outdoor scenes. They found that the best computational models perform similarly to people on outdoor images but most of their computational models perform worse on indoor images than on outdoor images. This was interpreted as indicating that the models used in the paper utilize more global information than local information, since indoor images have more objects and less global properties than do outdoor images.

In Vogel et. al, the authors tested multiple image manipulation methods and compared them to both a local and global computational model [36]. For their perceptual study, they asked people to classify scrambled, greyscale, and blurred images into the following five scene categories: coasts, rivers/lakes, forests, plains and mountains. They found that blurring and scrambling produce a similar drop overall in human performance but that different scene categories are affected in different ways. For example, rivers/lakes may require more global information to be correctly classified and so blurring affects performance less than scrambling.

The authors then created two computational models, one based on local information and the other based on global information, to determine which performs better and mimics human performance more. The semantic model, based on local information, divides images into 10x10 grids which are classified into one of nine concept classes. The frequency of each concept class is counted and stored in a concept occurence vector. A prototypical concept occurence vector is created for each of the five scene classes and each image is classified based on its minimum distance to each of the prototypical concept occurence vectors. For their global computational model, they utilize the computational gist model based on the work of Torralba et. al [33]. They found that the local semantic modelling approach performs better and nearly as well as human beings but their computational gist model is trained on 50 images per category, which may limit the strength of their conclusions.

In the third study, researchers compared a majority-vote computational model with people's performance on classifying degraded images [24]. Images were presented in one of three conditions: jumbled, jumbled with a random subset of blocks blacked out, and jumbled with blocks removed from the image (so that the effective size of each remaining block increased). People are better at classifying images in the 3rd condition then they are at classifying either jumbled images, or jumbled images with blacked out blocks, presumably because each block is enlarged in the 3rd condition so more information can be attained from the image. They then built a bag-of-words majority-vote computational model. This model describes each block in a jumbled image by computing the average RGB and HSV value in each 2x2 square of the block and recording these values to create a 120 dimensional feature vector. They then use k-means clustering to cluster each image in the training set to one of 500 code words. Each block in the test set is mapped to a code word and then they utilize a weighted majority-vote system to determine the class of the entire image. This model performed similarly to people on an outdoor dataset and worse than people on an indoor dataset, suggesting that their model could be an accurate representation of how people leverage local information. However, because images (both intact and jumbled) that are classified incorrectly by a large portion of people taking the survey are removed from analysis, it may not be true that the computational model exactly mirrors human performance. It may just be that the remaining images were less ambiguous.

While these three studies utilize computer models to determine if local or global information is important in recognizing the context of an image, they did not use neural networks for their computer vision algorithms. In the past few years, neural networks have revolutionized the field of computer vision and quickly overtaken other methods to become the current state-of-the-art for many difficult visual tasks. Because neural networks are very successful in image processing, we hope that by using neural networks to process distorted images, we will be able to learn more about how humans process information or how the human visual process system may be lacking (if the computer algorithm performs better than human beings). Neural networks were inspired by and share many similar properties as the human brain. It has been suggested that we can use neural networks to study how visual processing works in people [15] [7]. Some researchers have shown that convolutional neural networks can predict neuronal firing in inferior temporal cortex as well as the overall neuronal firing code [39]. Having a more robust and similar model of the human visual system will hopefully tell us more about the unique failures and successes of visual processing in people and how visual information is used to motivate specific visual tasks.

5 Overview of the Paper

In order to investigate what types of visual information are important for people and computers to perform scene recognition, we used various degradation methods to determine how removing certain image features affects scene classification. In Experiment One, we compare the accuracy of a pre-trained neural network and the GIST desciptor model with people on the indoor-outdoor scene identification task. We also train a neural network specifically for the indoor-outdoor scene recognition task. In Experiment Two, we reduce the outdoor bias of the neural network so that has a bias similar to people. In Experiment Three, we create and analyze a measure to compare local and global performance of each perceptual system on blurred and scrambled images. In Experiment Four, we train neural networks on degraded images to try and achieve similar classification results and local-and global-featural dependencies as people.

6 Methods

6.1 Image Selection and Manipulation

To ensure that the people in our perceptual study and computational models see a wide variety of images, it is necessary to pick a representative set of indoor and outdoor images that reflects the information and variance inherent in the two categories. To do this, we picked a diverse set of indoor and outdoor image categories from the Places dataset [40]. The Places dataset is a state-of-the-art image dataset with 205 scene categories containing nearly 2.5 million images. From this dataset, we chose 10 indoor and 10 outdoor categories and picked 10 random images from each of these categories in order to create a dataset of 200 images - 100 indoor and 100 outdoor images.

The indoor and outdoor scene classes are superordinate, meaning that they are almost mutually exclusive and exhaustive of all scene categories [34]. For indoor categories, we picked bedroom, classroom, dining room, kitchen, living room, lobby, museum, office, restaurant, and supermarket as our representative image categories. Our outdoor categories were coast, forest path, highway, mountain, skyscraper, valley, seacliff, river, residential neighborhood, and snowfield. These image categories capture a wide variance of the indoor and outdoor categories. The outdoor categories selected contain both outdoor natural scenes and outdoor manmade scenes. The indoor categories include a diverse representation of commonly observed indoor settings. These decisions were supported by others who looked at indoor and outdoor image classification tasks [16] [36] [34]. Each image is 256x256 pixels. Images were chosen from the validation set in the Places dataset, and are not used in training any of the neural networks we mention.

To process these images, we performed the following image degradation techniques: scrambling, blurring, grayscaling, adding noise, and reversing colors. An example outdoor image and its 35 different representations are shown in Figures 1-5.

For each image, we scrambled the pixels in blocksizes of 2, 4, 8, 16, 32, 64, and 128 pixels. The blocksize indicates the width and height of each square in the scrambled image. We broke up each image into blocks and then each block was mapped to a random location in the image. We scrambled images both with and without a grid marking each block of the image. For images with a grid, the gridline was chosen to be the average color of each image, in order to limit the effect of the grid on the image's average brightness and color. The grid was also added to the original, unscrambled images as a control condition to determine how much the presence of a grid in the image affects scene recognition for different perceptual systems.

To blur the image, we used a 2D Gaussian smooth-

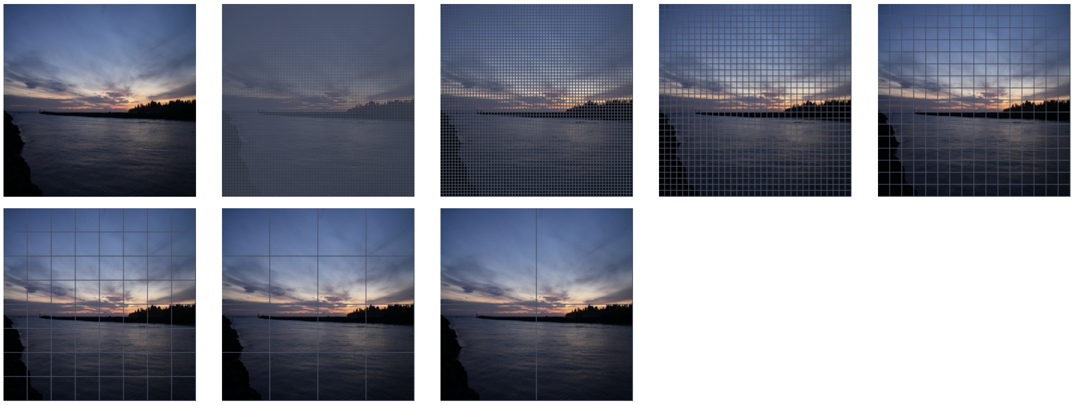


Figure 1: Intact images with a grid overlaid. The top-left image is the intact image without any manipulation. Grids of blocksizes of 2, 4, 8, 16, 32, 64, and 128 were added to the image. Grids were chosen to be the average color of the image.

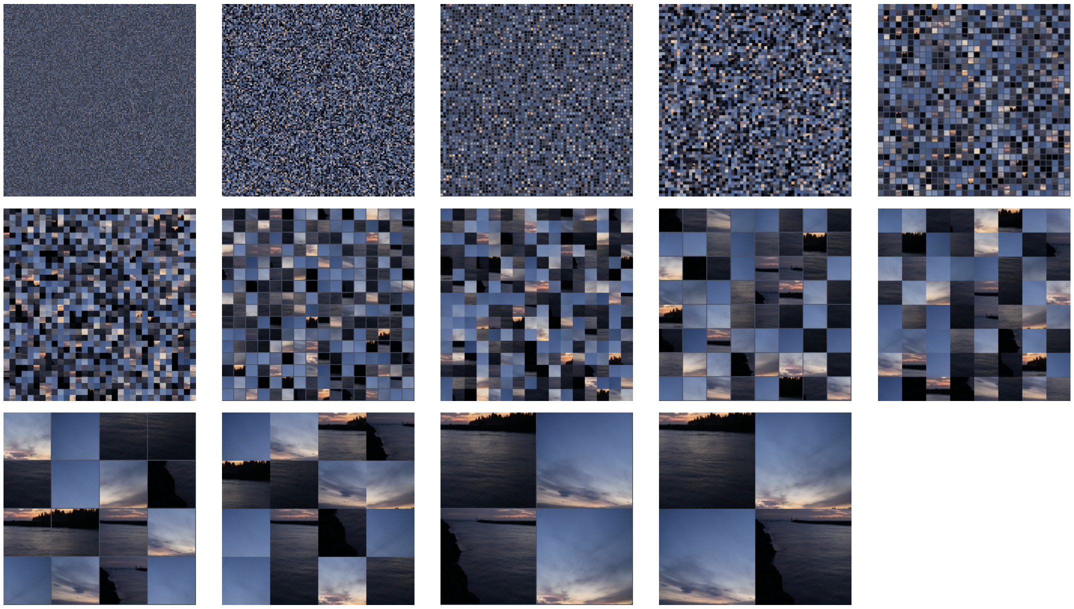


Figure 2: Scrambled images both with and without a grid. Blocksizes used are 2, 4, 8, 16, 32, 64, and 128 pixels.

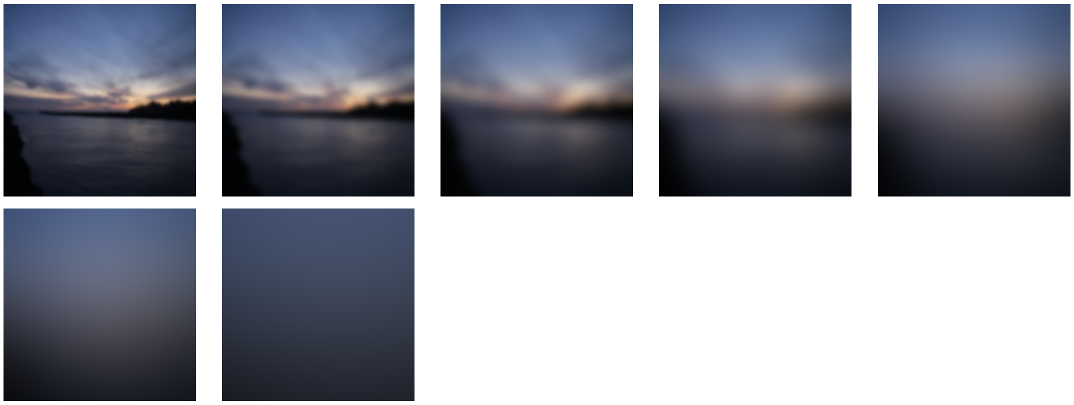


Figure 3: Degraded images blurred by a Gaussian smoothing kernel. Kernel with standard deviations of 2, 4, 8, 16, 32, 64, and 128. Images are shown in order of least blurry (standard deviation of 2) to most blurry (standard deviation of 128).

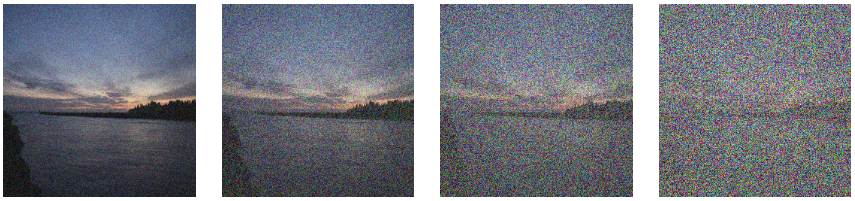


Figure 4: Degraded images with noised added. P values of 0.2, 0.4, 0.6, and 0.8 were used to determine how much noise to add to the image. Images are shown in order from least noise added to most noise added.

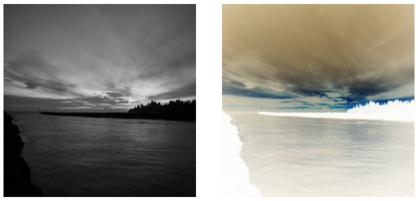


Figure 5: Degraded images by color. The image on the left is greyscale and the image on the right is the complementary color scheme of the intact image.

ing kernel [1]. For our experiments, we used kernels with standard deviations of 2, 4, 8, 16, 32, 64, and 128 pixels to math the blocksizes in the scrambled images. Replication padding was used.

To add noise, we created a uniformly distributed noise matrix with values drawn independently from 0 to 1 of size 256x256x3. We refactored the original image to have pixel values between 0 and 1. Then, we defined a percent variable p which determines the extent to which noise will be added to the image. For each pixel location (i, j) in the new image, the pixel value is determined by Equation 1. We used values of 0.2, 0.4, 0.6, and 0.8 for p to create 4 noisy images. nim, im, and r denote the new image, original image, and noise image, respectively.

$$\mathrm{nim}(i, j) = (1 - p) * \mathrm{im}(i, j) + p * \mathrm{r}(i, j) \tag{1}$$

For greyscale degradation, we converted each RGB image into a simple greyscale image. For each pixel value in an image, we created the new greyscale image by taking the R, G, and B pixel values for pixel location (i, j) in the image and computing the new value by Equation 2. For creating the complement image, we subtracted each pixel value from 255, the maximum pixel intensity (see 3). The new image becomes the color complement of the original image - white pixels become black, reddish pixels become green, and so on. These manipulations were used to measure the importance of color on the classification task.

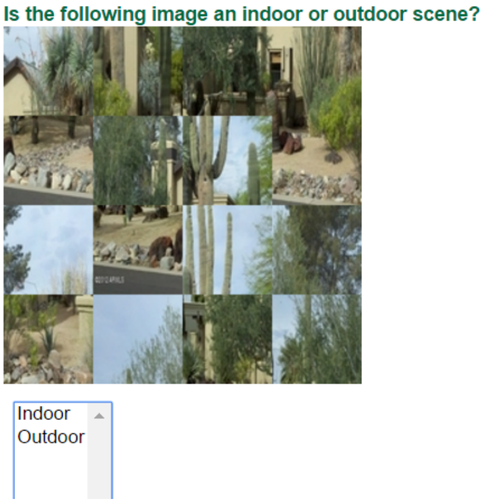
$$\begin{aligned}\mathrm{nim}(i, j) = 0.2989 * R(i, j) + 0.5870 * G(i, j) \\ +0.1140 * B(i, j)\end{aligned} \tag{2}$$

$$\mathrm{nim} = 255 - \mathrm{im} \tag{3}$$

After these manipulations, each of the 200 images had 35 versions, 14 of which were scrambled versions with and with out a grid, 7 of which were blurred images, 4 of which were noisy images, 2 for greyscale and complementary color patterns, 7 for the original image with a grid, and 1 for the original image without any degradations.

6.2 Perceptual Study

To collect human scene classification performance on the test set, we created a Qualtrics survey which we posted to Amazon Mechanical Turk. The survey showed 204 images. 200 images were from the test set shown in a random order. Each image was shown exactly once and its exact occurence (whether it was blurred, scrambled, etc.) was chosen randomly. Four images were included as catch trials. These images were easily identifiable, intact images. Participants who missed a catch trial had their results discarded from the analysis. Participants were paid $2.75 for their participation. We obtained results from 222 people and threw out results from 16 people either for missing one of the catch trials or completing the survey multiple times. In total, there were 206 utilizable responses. An example question is shown in Figure 6. The study procedure was approved by the institutional review board at Dartmouth College.



Is the following image an indoor or outdoor scene?

Figure 6: Depiction of the question text.

6.3 GIST Classifier

To compare the results collected from the perceptual study to computational approaches, we use two types of models: a GIST descriptor model and a convolutional neural network. Using Oliva et. al's feature extractor for the GIST description, we trained a support vector machine and a linear discriminant classifier using training images from the Places dataset [22]. Linear discriminant analysis (LDA) computes a linear decision boundary between the indoor and outdoor image classes. For this reason, it may not be as accurate as other classifiers which can compute nonlinear decision boundaries. It generally is easier to train and much more computationally efficient then

[1] In Experiment Three, we discuss another blurring technique used to compare classification performance on blurring and scrambling degradation techniques. For Experiments 1-2, we use the blurring method discussed here.

non-linear classifiers but may not achieve the same accuracy. It is also less prone to overfitting the training set since there are not enough parameters (one per feature) to learn the noise in the data. Support vector machines (SVMs) can compute non-linear decision boundaries by increasing the dimensionality of the input data and then computing a linear decision boundary on this new data. For this reason, they can be more accurate than simple linear decision boundaries.

For our purposes, we trained each classifier on 250 random imges from each of the 205 image categories for a total of 51250 training images. We computed a GIST descriptor vector (with 512 features) for each image in the training set and used the image's indoor or outdoor category as its label. We also computed a GIST descriptor for each of the 7000 images in the image test set in order to compute classification accuracies for these LDA and SVM GIST models.

6.4 Neural Network

Table 1: This table shows the neural network architecture used throughout the paper. Each convolutional and fully connected layer is followed by a rectifying nonlinearity. The final layer contains 205 neurons and is used to compute the class probabilities, using a softmax objective function.

Layer Name	Size
Data	227x227x3
Conv1	96 11x11 filters of stride 4, pad 0
Norm1	
Pool1	3x3 filters of stride 2
Conv2	256 5x5 filters of stride 1, pad 2
Norm2	
Pool2	3x3 filters of stride 2
Conv3	384 3x3 filters of stride 1, pad 1
Conv4	384 3x3 filters of stride 1, pad 1
Conv5	256 3x3 filters at stride 1, pad 1
Pool3	3x3 filters of stride 2
FC6	4096 neurons
FC7	4096 neurons
FC8	205 neurons (class scores)

We used a pretrained deep convolutional neural network, Places205-AlexNet, which is a version of AlexNet trained on the 205 scene categories in the Places database. Places205-AlexNet's architecture is described in Table 1. The 256x256 images are downscaled to 227x227. The input is then fed through a series of 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 so 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 3 fully connected layers; he last of which computes class scores for 1000 classes (in the case of AlexNet). For Places205-AlexNet, the final layer is a fully connected layer with 205 neurons that produce a 205-dimensional vector containing the class scores for the 205 classes in the Places database.

We implemented this neural network architecture with both 205 classes and two classes corresponding to just indoor and outdoor images. For the network with a final layer of 205 outputs, we used the authors' [40] mapping of each scene category to indoor/outdoor to get the true label for the image. We used both top-1 and top-5 categorization for the network trained on 205 scene classes to determine the indoor or outdoor class of the image. For top-5 categorization, we took the 5 most likely scene classes and counted how many were outdoor or indoor. The true label was then taken to be the class (outdoor/indoor) with the most representatives in the top-5. For top-1 classification, we took the most likely scene class, converted it to an indoor/outdoor label, and returned this value as the true label. For neural networks trained with a final layer of 2 outputs, the true label is taken to be the most likely class (with no additional computations).

Training and testing are implemented through the Caffe neural network library. All neural networks are trained with the parameters shown in Table 2. A softmax loss layer was added to the neural network to compute the learning objective. The results from different training strategies are discussed below.

Table 2: Neural network training parameters

Parameter	Value
Regularization	L2
Weight Decay	0.0005
Momentum	0.9
Learning Rate Strategy	step
Base Learning Rate	0.01
Gamma	0.1
Step size	100000

6.5 Calculating Bias

In addition to computing classification accuracy, we also calculated d-prime and criterion statistics to get a measure of how biased a perceptual system is. All calculations are done with respect to indoor images as the signal and outdoor images as the noise. We used the hits, misses, false alarms, and correct rejections to calculate these statistics. A hit is defined as a correct classifcation of an indoor image. If a system classifies an indoor image as an outdoor image, then that counts as a miss. Conversely, a false alarm occurs when an outdoor image is classified as an indoor image. A correct rejection means the system classified an outdoor image correctly. Using this terminology, we calculated d-prime and criterion as follows.

The hit rate (hr) and false alarm rate (fa rate) are calculated as in Equations 4 and 5, where fas denoted false alarms and crs denotes correct rejections. Then d-prime and criterion are calculated based on Equations 6 and 7. The function Z represents the the inverse of the cumulative distribution function of the Gaussian distribution.

$$\text{hr} = \frac{\text{hits}}{\text{hits} + \text{misses}} \tag{4}$$

$$\text{fa rate} = \frac{\text{fas}}{\text{fas} + \text{crs}} \tag{5}$$

$$\text{d-prime} = Z(\text{hr}) - Z(\text{fa rate}) \tag{6}$$

$$\text{criterion} = -\frac{Z(\text{hr}) + Z(\text{fa rate})}{2} \tag{7}$$

The d-prime index measures how sensitive the system is to noise. A higher d-prime indicates that the signal in the images can be more readily detected. The criterion index provides a measure of how biased the system is to the two image classes. A high negative criterion value indicates a high rate of both false alarms and hits, which suggests the system is only picking one class. In contrast, a high positive criterion value indicates that the system is more prone to picking the opposite class (in this case outdoor). A near-zero criterion value implies that there is no strong bias in the system.

7 Experiment One - Perceptual Study and Computational Performance

In this section, we present the classification results from the perceptual study, the pre-trained neural network, and the SVM GIST classifier on the test set. The SVM GIST classifier proved to be better than the LDA classifier on our test set so we omit the LDA classifier from our analysis (the plots for the LDA are shown in the appendix). We show plots of percent correct, d-prime and criterion statistics. The d-prime and criterion statistics are plotted with respect to the indoor images as the signal and the outdoor images as the noise.

Section 7.1 analyzes each of the three models' performances on different degradation techniques. Section 7.2 contains the results from training AlexNet for the task we are interested in (indoor versus outdoor image classification). This is done by replacing the final layer of 205 units with a layer that contains 2 units and running the training on the Places205-database, with each of the 205 classes converted to an indoor or outdoor label.

7.1 Overall Performance

The pre-trained neural network's top-5 performance on the original images is 98.5%. This is slightly less than human performance but better than the SVM model, which achieves an accuracy of 85%. The neural network and people achieve similar classification accuracies for low-level degradations but the neural network seems to be negatively affected by manipulations more than people are.

Figure 7 shows performance of each perceptual system on various degradation techniques. Figures 8 and 9 display the d-prime and criterion statistics respectively for these systems. Points farther to the right on the x-axis indicate stronger degradation levels.

Panel 7.A shows the performance on intact images with a grid overlaid. While people's classification accuracy is relatively stable no matter the blocksize of the grid on the image, the computational models experience a dropoff as the blocksizes get smaller and smaller. Most noticeably, once blocksizes reach around 16x16 pixels, both the neural network's and the SVM's performances drop. The d-prime and criterion statistics (in Panels 8.A and 9.A) indicate that most visual systems are biased to classifying these degraded images as outdoor, with people being the

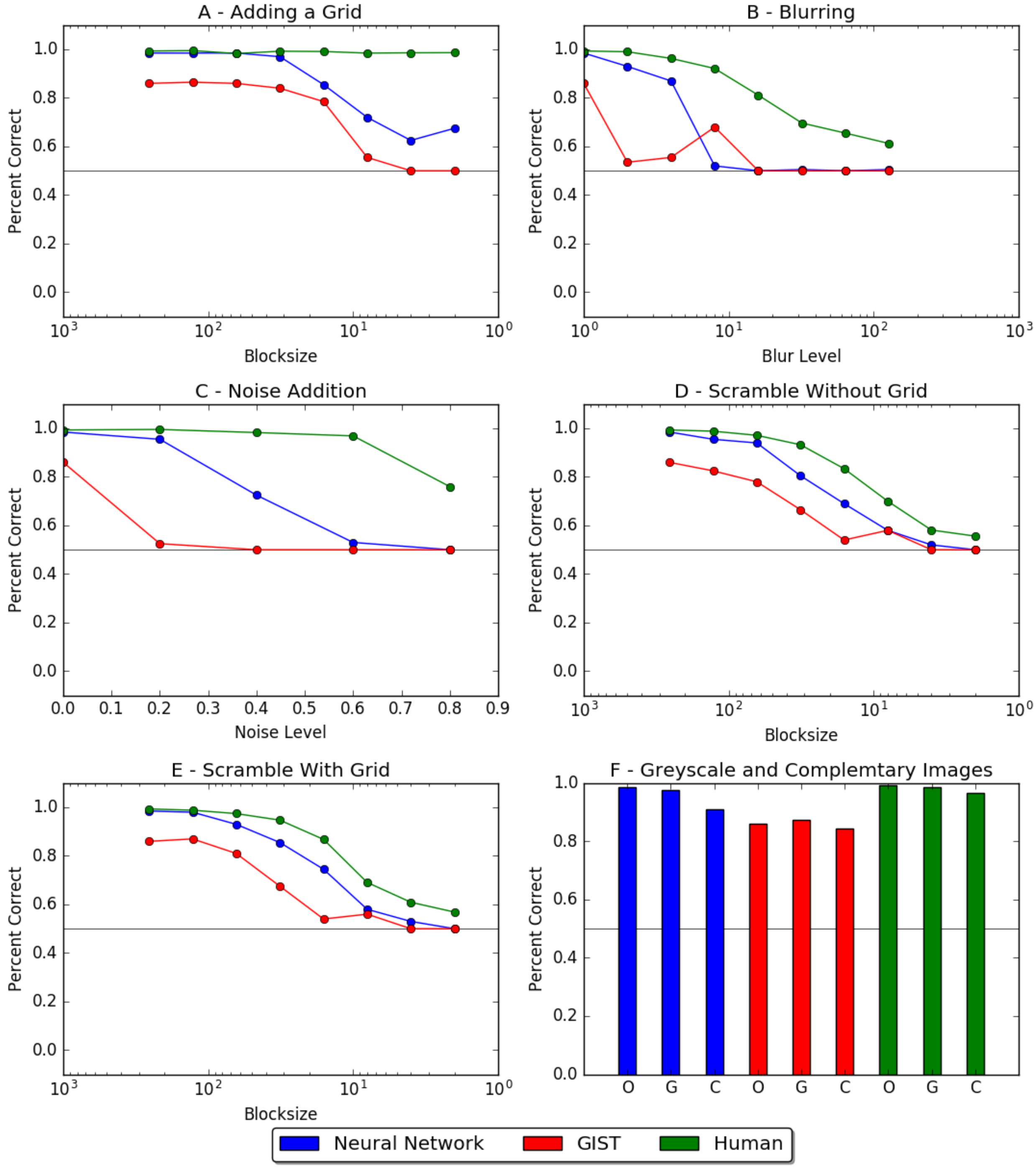


Figure 7: Percent correct by image manipulation for several perceptual systems - a neural network, a GIST SVM model, and human subjects. The x-axis is oriented such that low-level/no manipulations are on the left and higher-level manipulations are on the right. In Panel F, O stands for original images, G for greyscale images, and C for complementary images. The black horizontal line in each plot represents chance.

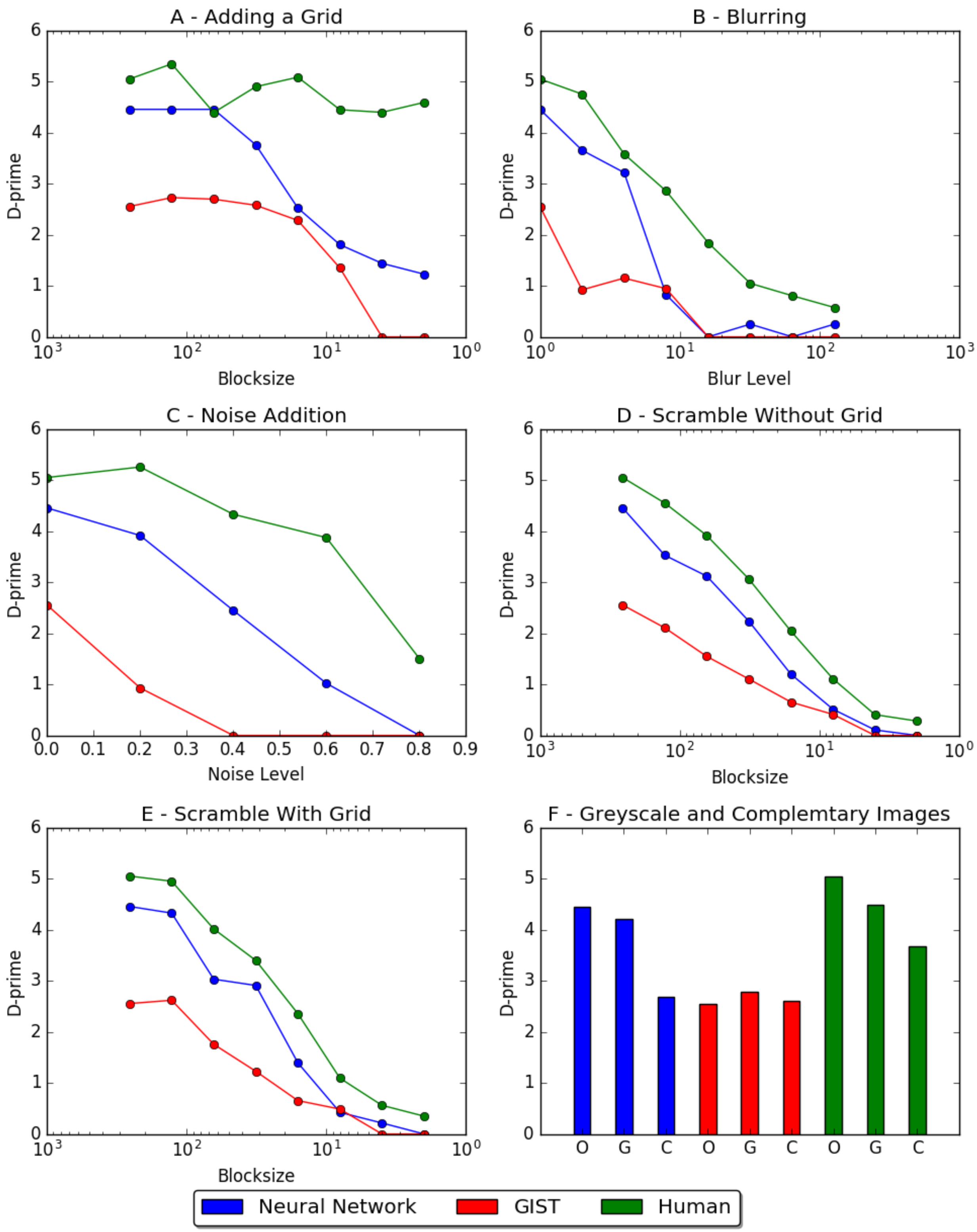


Figure 8: D-prime statistics for the neural network model, GIST SVM model, and human subjects on various image manipulations. Values closer to 0 indicate a low sensitivity, whereas higher values indicate high sensitivity.

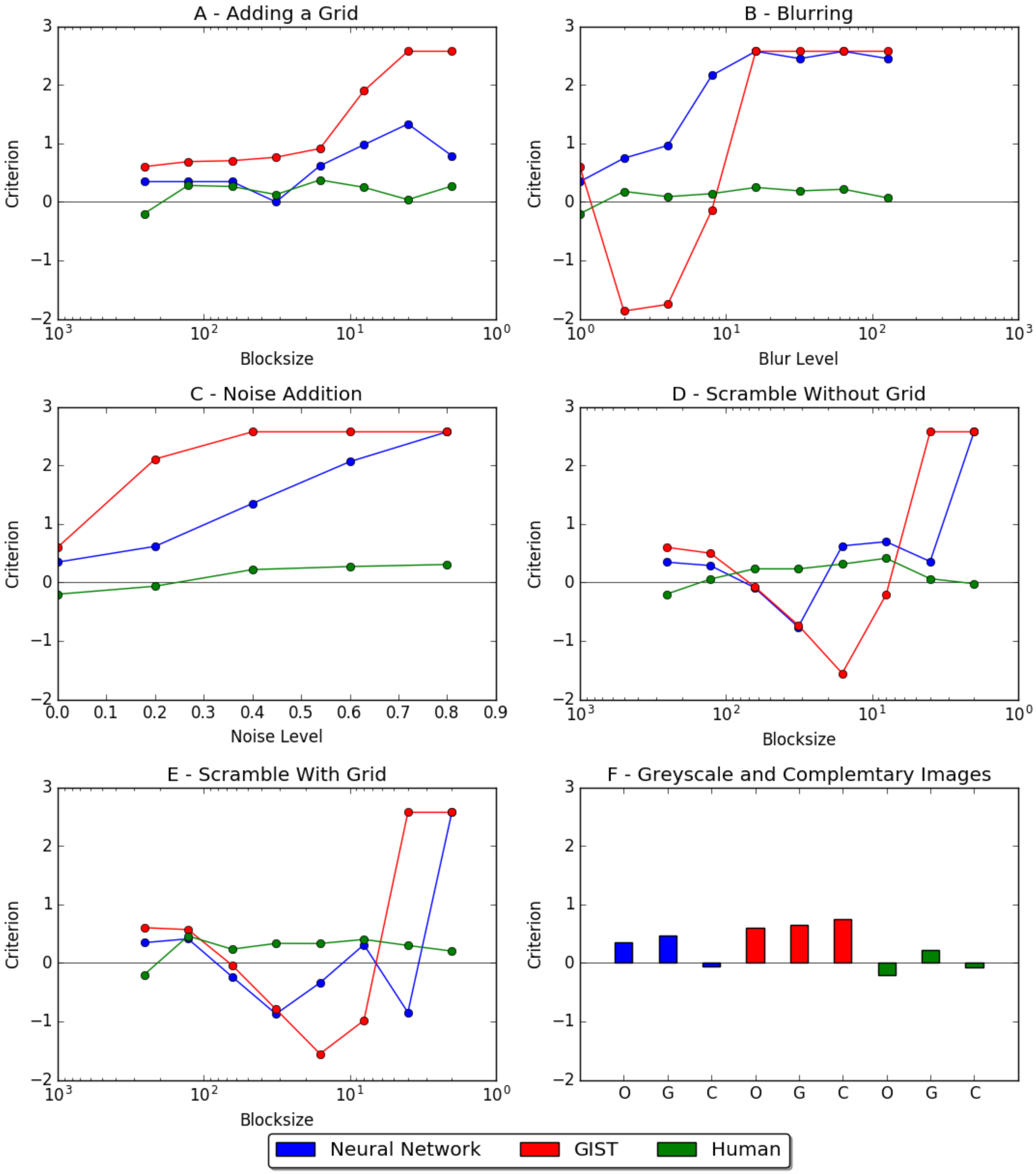


Figure 9: Criterion statistics for the neural network, GIST SVM model, and human subjects on various image manipulations. High positive values indicate an outdoor bias whereas high negative values indicate an indoor bias. The black horizontal line indicates no bias (value 0).

least biased and the SVM gist model being the most biased.

For blurry images, Panel 7.B shows the classification results, with the left most point being intact, unblurred images and the right most point being the highest blur level. As expected, performance drops as the blur level increases. Suprisingly, however, both computational models' classification accuracies reach chance very quickly. On the contrary, human subjects classify all levels of blurred images above chance. It is interesting to see that the bias of the computational models toward outdoor images is very apparent with blurred images, with almost all images being classified as outdoors when blur levels are high (Panel 9.B). For the SVM GIST model, at mid-level blurring, the model has an indoor bias but at higher-blur levels, it has the same bias as the neural network.

Panel 7.C shows the performance for noisy images. Most of the same trends can be observed here as with blurry images. The SVM GIST model's classification accuracy drops off very quickly as images become more degraded. The neural network maintains its near-human level performance but as degradations get stronger, its performance also approaches chance. The same bias toward outdoor images can be seen in Panel 9.C. People's performance only shows a substantial decrease for the most noisy images, which indicates that humans are remarkable at interpreting noisy images.

The results for scrambled images are displayed in Panels 7.D and 7.E. The neural network has a similar performance to people until blocksizes get lower than 64 pixels when the neural network's performance drops below human performance. At smaller blocksizes, less than 4 pixels, the neural network, the SVM GIST model, and human subjects all perform at roughly chance levels. However, people are only slightly biased toward outdoor images while the neural network and support vector machine are very biased toward outdoor images (Panels 9.D and 9.E). It is also noticeable that at medium scramble levels, the neural network and SVM are biased toward indoor images.

The effect of adding a grid to scrambled images can be measured by comparing the two plots. For people, it seems that scrambling an image without a grid increases performance. The grid could be hiding pixels that are important for classification, which could account for this slight decrease in performance. The grid makes color information disappear (such as the variance in color in the image since the grid is the average color), especially at smaller blocksizes, so the classification task becomes harder for people. Likewise, for both computational models, the grid negatively impacts classification performance. By hiding relevant pixel information and creating sharp edges, there is less useful image information on which to ground classification. This results in slightly decreased performance with a grid.

Panel 7.F displays the results from the color manipulations. All perceptual systems experience a slight drop in performance with complementary images, but almost no effect on classification accuracy with greyscale images. It is also clear that the neural network and GIST SVM model are slightly biased toward outdoor images when classifying complementary images (see Panel 9.F). Although there is a slight bias for classifying images as outdoor, it is not nearly as strong as for the other degradation techniques. The neural network also has a very slight bias toward indoor, complementary images.

7.2 Retraining AlexNet on the Indoor/Outdoor Image Classification Task

Next, we removed the final fully connected layer with 205 units of AlexNet and replaced it with a fully connected layer with 2 units in order to build a neural network trained specifically for the indoor versus outdoor classification task. The other parts of the AlexNet architecture were not changed. We ran training from scratch, using the same training set and validation set as AlexNet trained on the Places205 database. The data was relabelled to correspond to the indoor versus outdoor classification task. The network was trained for 350000 iterations.

Table 3 shows a comparison between the pre-trained neural network and our neural network trained at the indoor versus outdoor classification task. The number next to the manipulation technique indicates the blocksize in the case of original and scrambled, the standard deviation for the Gaussian filter in the case of blurred images, and the noise level for the noisy images. The neural networks achieved similar accuracies on most manipulations, with the pre-trained neural network marginally outperforming the indoor versus outdoor neural network for some degradations. Plots that display the percent correct, criterion, and d-prime statistics for people, the pre-trained neural network, and the indoor/outdoor neural network can be seen in the Appendix in

Figures 22 - 24. These figures demonstrate that both types of neural networks achieve similar classification performance and have similar biases toward outdoor images.

8 Experiment Two - Attempting to Eliminate the Neural Network Bias

One problem with the Places database for the task of indoor/outdoor scene recognition is that there are more outdoor scene categories than indoor scene categories. Out of the 205 scene categories, 69 of them are indoor and 136 of them are outdoor. This means that the neural network sees more outdoor images than indoor images while being trained, which may account for the bias toward classifying scenes as outdoors. Another possibility is that because we use the top-5 scene classes to determine if an image is indoors or outdoors, it may just be the case that the neural network has a higher baseline probability of picking an outdoor scene. Since there are more outdoor scene categories, if we look at the top-5 categories for each image, then it is expected that there will be more outdoor scene categories.

To remedy this issue, we tried a number of potential solutions. The first included using top-1 rather than top-5 accuracy to determine if an image is indoor or outdoor. We converted each of the 205 scene categories to either indoor or outdoor (using a mapping provided by the authors of the Places database) and selected the scene category with the highest probability. Table 4 shows the percent accuracies for top-5 and top-1 results using the pre-trained Places205-AlexNet. Overall, a drop in performance is noticed when we move to top-1 classification. It is still true that performance on heavily distorted images is around 50%. Performance for heavily scrambled images falls below 50%. This indicates that the network is not exclusively classifying heavily scrambled images as outdoors (which was the case with the top-5 network), since it is getting some of the outdoor images wrong.

We also looked at the criterion statistics for the top-1 neural network. Figure 10 reveals that the bias for outdoor images still exists. The criterion for the top-1 neural network is a few points lower than the 2-class neural network but the behavior and shape of the curve is the same. For scrambled images, however, the criterion drop belows 0, indicating a bias for indoor images with large blocksizes. This mirrors the behavior of the SVM computational gist model. This could be because the neural network needs to be biased to predict an indoor scene category in order to have high top-5 classification accuracy on indoor vs. outdoor tasks, but it still selects more outdoor categories in general. However, it does not seem that we've made a more indoor-biased neural network since at high blur, noise, and scramble levels, there is still a tendency to classify all images as outdoors.

The second potential solution was to train a neural network using an equal number of indoor and outdoor images. Out of the 2,448,872 images in the Places database, 728,143 of them are indoor. We trained a new neural network on the indoor/outdoor classification task (2 units in the final fully connected layer) using a training set with 728,143 indoor and outdoor images for a total of 1,456,286 images. We did this by selecting all images in indoor scene categories and grouping all outdoor images in a list and randomly picking the first 728,143 images. If there is not an inherent bias with the neural network architecture toward outdoor images then this change should produce a less biased classification model. However, it may be that heavily degraded images have more in common with image features in outdoor classes (e.g. one overarching color, grainy texture, etc.), and this makes it hard to detect indoor features in a degraded image.

After running the newly trained neural network on the test set, we obtained the criterion results shown in Figure 10 (Figure 11 compares the classification accuracy of this neural network with results collected from the perceptual study and the network trained in Experiment One). While the neural network trained on an equal number of indoor and outdoor images has the same outdoor bias for blurred images (the criterion is positive and follows the criterion line for the original network), it seems that the reverse is true for scrambled images at middle-range blocksizes. The criterion becomes negative for mid-level distortions. However, the neural network has as strong an outdoor bias as the original networks for very scrambled images (2 pixel blocksizes). For noisy and original images with a grid, this type of neural network has less bias than the top-1 and original neural network. Although, the network seems less biased toward outdoor images, it still does not mimic the behavior of people, who have very little bias for either indoor or outdoor images.

Looking at the Figure 11, it is also interesting to

Table 3: Pre-trained Neural Network (Places205-AlexNet) vs. Our Trained Neural Network on the indoor/outdoor classification task. Each value is a percentage of images answered correctly for the various types of image manipulations and levels of degradation.

	Pre-trained NN	Our NN		Pre-trained NN	Our NN
Original	98.5	97.5	Blurred-2	93	95
Grid-2	67.5	82	Blurred-4	87	92.5
Grid-4	62.5	69	Blurred-8	52	64.5
Grid-8	72	78	Blurred-16	50	50
Grid-16	85.5	90.5	Blurred-32	50.5	50
Grid-32	97	96	Blurred-64	50	50
Grid-64	98.5	96.5	Blurred-128	50.5	50
Grid-128	98.5	97	Noisy-0.2	95.5	92.5
Scrambled-2	50	49.5	Noisy-0.4	72.5	65.5
Scrambled-4	53	50	Noisy-0.6	53	50
Scrambled-8	58	71.5	Noisy-0.8	50	49.5
Scrambled-16	74.5	77	Grayscale	97.5	95.5
Scrambled-32	85.5	80.5	Complement	91	87
Scrambled-64	93	92.5			
Scrambled-128	98	96.5			

Table 4: Top 5 pre-trained neural network vs. top 1 pre-trained neural network. Each value is a percentage of images answered correctly.

	Top 5	Top 1		Top 5	Top 1
Original	98.5	96	Blurred-2	93	93.5
Original-2	67.5	70	Blurred-4	87	89.5
Original-4	62.5	62	Blurred-8	52	50.5
Original-8	72	74.5	Blurred-16	50	50
Original-16	85.5	79.5	Blurred-32	50.5	50
Original-32	97	86.5	Blurred-64	50	50
Original-64	98.5	95.5	Blurred-128	50.5	50
Original-128	98.5	97.5	Noisy-0.2	95.5	92.5
Scrambled-2	50	57	Noisy-0.4	72.5	73.5
Scrambled-4	53	67	Noisy-0.6	53	59.5
Scrambled-8	58	75.5	Noisy-0.8	50	50.5
Scrambled-16	74.5	72.5	Grayscale	97.5	97
Scrambled-32	85.5	81	Complement	91	80.5
Scrambled-64	93	90.5			
Scrambled-128	98	95			

note that this neural network performs more similarly to people for some manipulation techniques. Adding a grid to an image has a much smaller effect on classification accuracy than it did on the previous neural networks (Panel 11.A). Similarly, blurring and adding noise to an image do not have as big an effect on classification accuracy. However, scrambling an image seems to keep performance the same, or worsen it at some blocksizes. This suggests that reducing certain biases in the neural network will bring classification accuracy of degraded images closer to people.

In order to obtain a result that more closely resembles people, it is necessary to figure out how to train out this bias. Although humans are slightly biased toward outdoor images as well (humans classified 22,033 images as outdoor vs. 19,367 images as indoor in the Qualtrics survey), they are not prone to the same level of biases as computational models. A summary of criterion statistics can be seen in Figure 12. Although we have reduced the outdoor bias with the neural network trained on an equal number of indoor and outdoor images, it is still obvious that humans are the least biased and other systems may still be biased in the other direction (toward indoor images), a behavior not observed in the perceptual study.

9 Experiment Three - Local vs. Global Computations

In order to determine if a perceptual system utilizes more local or global information, we created a new filter that allows us to compare performance on blurred and scrambled images. Blurred images degrade local information, while scrambled images degrade global information. If we can match each level of degradation for both blurred and scrambled images, then we can come up with a systematic way to compare performance on these two image manipulations. We then reran the perceptual study using this new blurring filter and collected results from our computational models to determine how big of an impact local and global information play in scene recognition for different perceptual systems.

9.1 Methods

For the new blur filter, we created a mapping between the blocksizes in the scrambled images to blur levels in the blurry images. If an image is scrambled into a blocksize x, then the maximum spatial frequency s_f (in cycles/block) is computed as in Equation 8.

$$s_f = x/2 \tag{8}$$

Hence, the entire contiguous spatial frequency of the scrambled image with blocksize x is no more than s_f, the maximal block frequency. In order to represent blurred images along this scale, we created a 5-th order low pass Butterworth filter to filter images in the frequency domain. To filter images, we broke each image into its red, green and blue image components. Then, we took the Fourier transform of each color component to compute the frequency domain of the image. We convolved the frequency domain of the image with a 5-th order Butterworth filter. Finally, we converted the image back into the spatial domain to create a blurred image.

Examples of one set of blurred images (in order of most blurry to least blurry) produced by this process are shown in Figure 13. Frequency cutoffs were chosen to correspond to the blocksizes in the scrambled images from the above equation (but in units of cycles/image). This way, we can compare classification accuracy for blurred images with accuracy for scrambled images in order to determine how local or global features affect different perceptual systems. Figure 25 in the Appendix illustrate the filters in the Fourier domain for each of the frequency cutoffs.

If a perceptual system utilizes more global features in performing classification, then it is expected that it would classify blurred images, which possess more global information, more accurately than scrambled images since the spatial, global relations of the scene remain intact. If, on the contrary, a perceptual system is heavily affected by local features, than it should be able to classify more scrambled images correctly than blurred images, since scrambled images maintain local information such as properties of the objects in the image.

To determine if a perceptual system is more local or global we define a Difference Index between classification performance on scrambled and blurred images. For a given classification accuracy on blurred images b and an accuracy on scrambled images s, we define the Difference Index as follows:

$$\text{Difference Index} = b - s \tag{9}$$

We expect more global visual systems to have Difference Index values above 0 and local perceptual systems to have Difference Index values below 0.

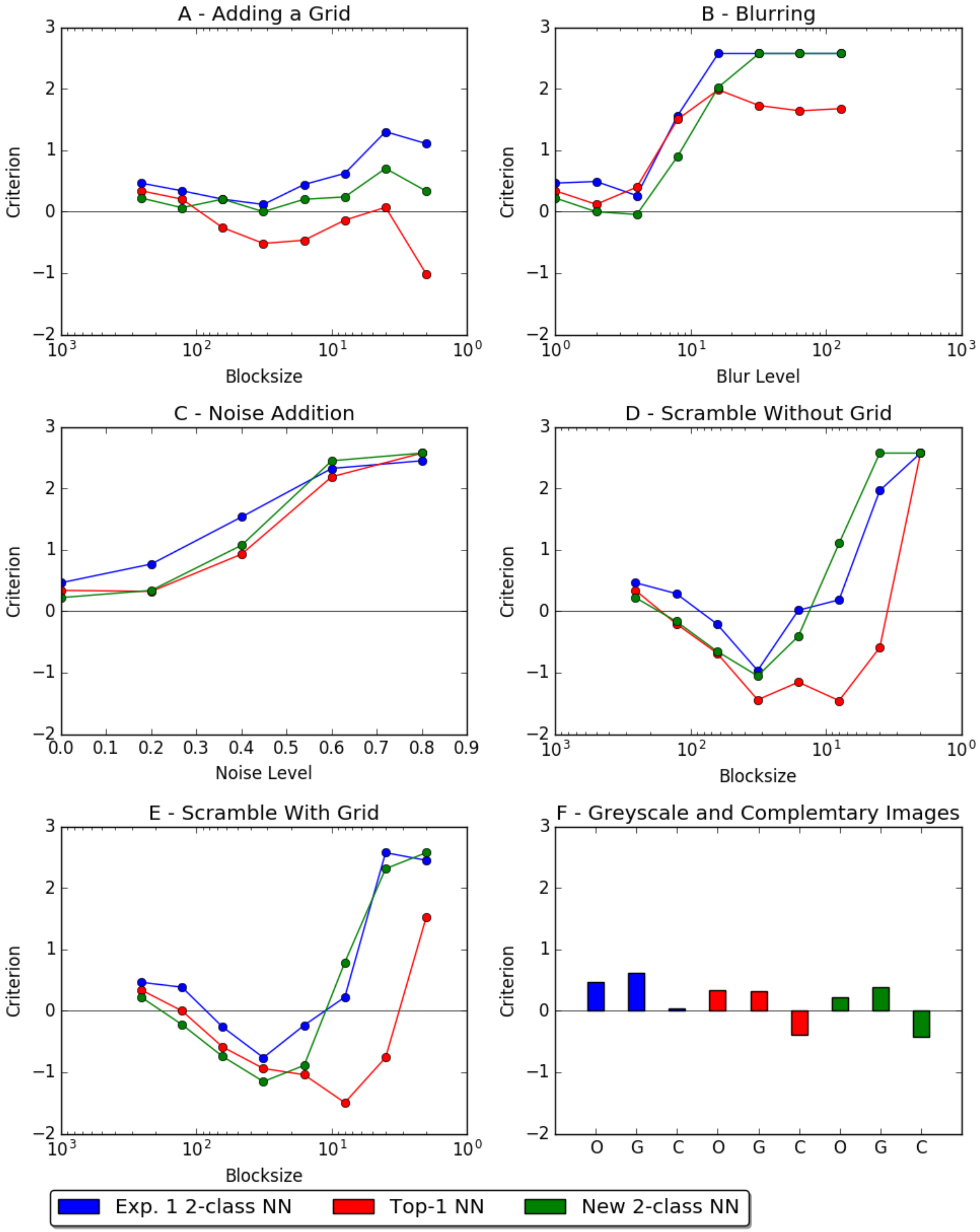


Figure 10: Comparison of the biases of three neural networks. The blue neural network is the Indoor/Outdoor network from Experiment one, the green network is the top-1 205 class neural network, and the red line indicates the new Indoor/Outdoor neural network trained on an equal number of indoor and outdoor images.

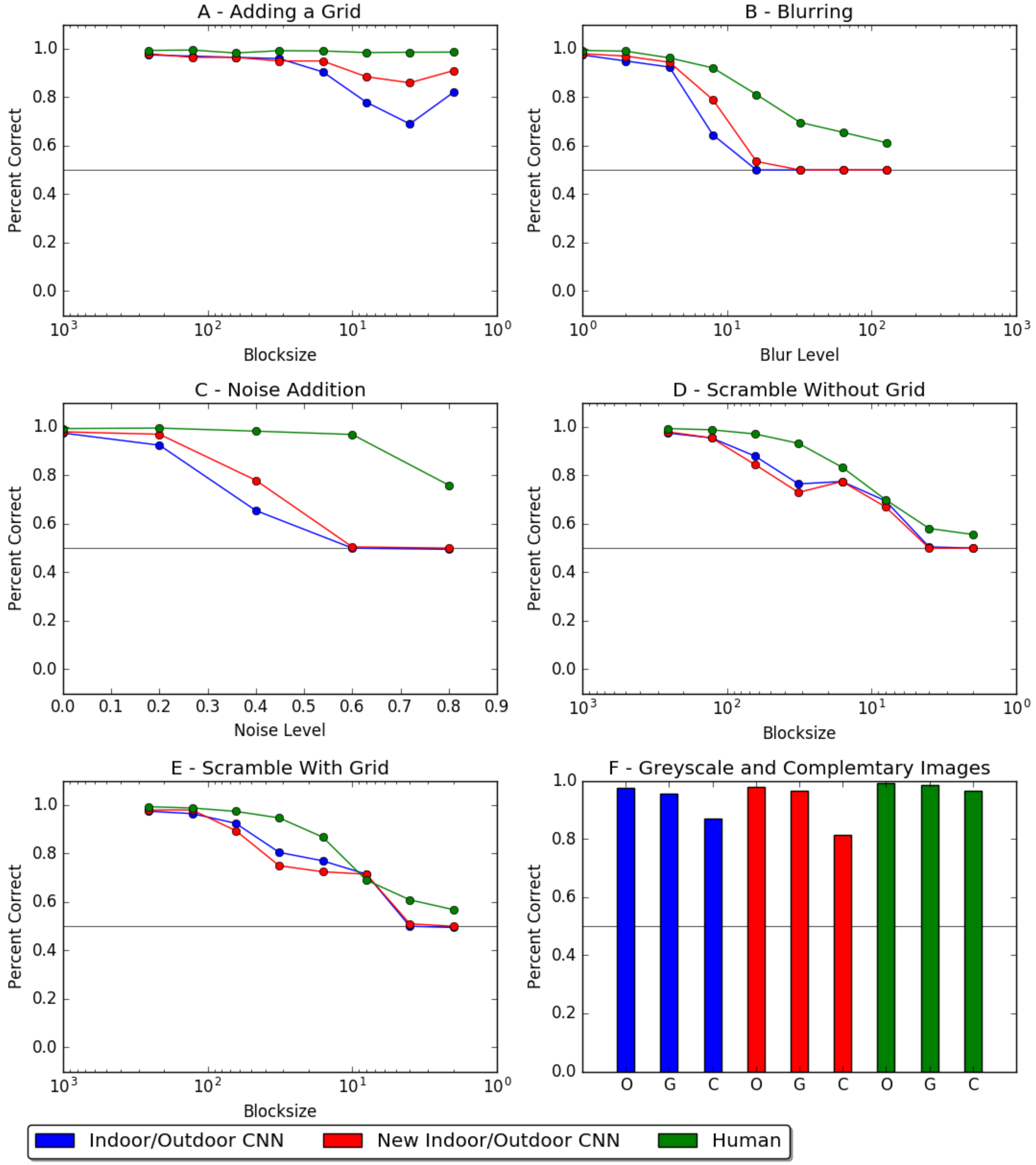


Figure 11: Plot of human results and two neural networks - the Indoor/Outdoor network from Experiment One, and a new neural network trained on an equal number of indoor and outdoor images.

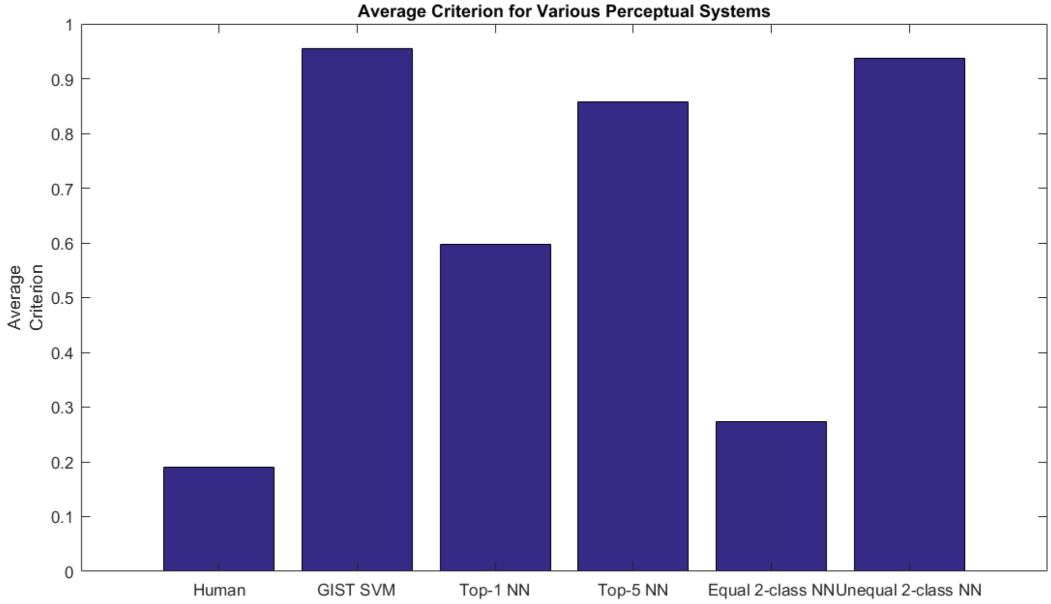


Figure 12: Average Criterion Values for people and various neural networks. People have the lowest bias toward outdoor images, followed by the neural network trained in Experiment Two on an equal number of indoor and outdoor images.

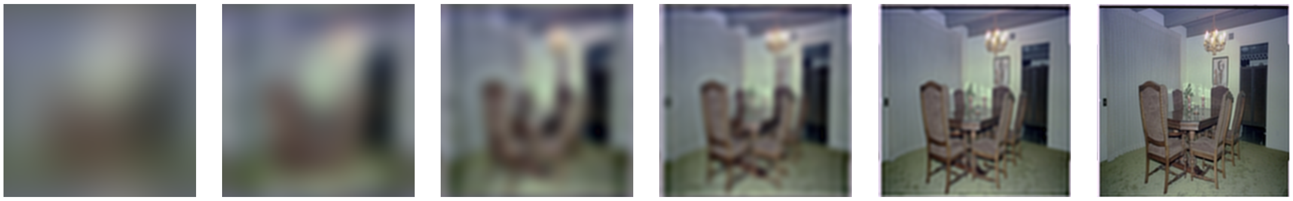


Figure 13: 6 exemplars of each blurred image (blurred by a 5-th order Butterworth filter with cutoff frequencies of 2, 4, 8, 16, 32, and 64 cycles/image. The right-most image has the highest cutoff frequency and left-most has the lowest cutoff frequency.

To test people on these new blurred images, we conducted a second perceptual study. This study had the same setup, questions and test set as the first study except we tested people on 13 manipulations for each image, rather than 35: 6 for the new blurred images (with frequency cutoffs set as 2, 4, 8, 16, 32, and 64), 6 scrambled images (without a grid, with blocksizes of 4, 8, 16, 32, 64, and 128), and the original, intact images. The survey showed 204 images, 4 catch trials and 200 images from the test set. We analyzed results from 170 participants.

9.2 Results

The results (using the Difference Index) for the human visual system are shown in Figure 14.A. The top panel shows the accuracy values for classifying both scrambled and blurred images while the bottom panel shows the Difference Index at each degradation level. While low-level degradations produce little-to-no difference in local or global image performance, mid-level and higher distortions seem to favor global visual processing. Eliminating more spatial frequencies in blurred images and having small blocksizes results in slightly better performance for blurred images, indicating that global features are more important in determining the image's class. These results suggest that at low-level manipulations, both local and global image properties play a role in determining the contextual features of the image. However, as local and global information are both heavily degraded, humans can more accurately classify images with global image properties than with local image properties. This suggests that global features may be used more by the human visual system in scene recognition, especially when most image information is unavailable.

The neural network's local versus global information is shown in Column 14.B. The neural network relies mostly on global information for mid-level degradations as it achieves higher classification accuracies on blurred images than scrambled images. Its Difference Index is positive for most degradations, indicating the neural network is more negatively affected by global manipulations than local. However, at the second most degraded point, the neural network's Difference Index falls below 0, meaning that the network is less robust to blurred images, a local image manipulation. This makes sense because a convolutional neural network looks at patches in the image and learns features from those patches. As long as there are patches in the image that are big enough to possess pertinent visual features, which scrambled images (with a blocksize of 8 pixels) still have, then the convolutional network will be able to determine the features in the patch. The most degraded point (blocksizes of 4 pixels) may posses less visual information in each patch than do blurred images at the same degradation level, which could account for why the network's Difference Index is positive at that point. There seems to be a point at which removing more global information reduces performance then the equivalent removal of local information.

One similarity between the neural network model and the human visual system is that both systems start off responding more to global information than to local information. With large blocksizes and high frequency cutoffs, both systems are better at classifying blurred images. This indicates that the overall features of the scene, such as openness and depth, are more useful than the local, object-based features of the scene. When patches get so small that local object information becomes harder to identify, then people have a strong preference for global features, which suggests that people are more capable of identifying the context of the scene just from global features, but they have an easier time when there are also local features. The neural network, however, is more affected by local manipulations (signalling a negative Difference Index) at high levels of image degradation.

The SVM GIST model (seen in Column 14.C) has both positive and negative Difference Index values depending on the degradation level. At low-levels of degradation, the GIST model is more affected by global manipulations than local. This is intuitive since the GIST model is created to compute feature vectors that relate global relationships in the image. However, at most other points, the GIST model can more accurately classify scrambled images than blurred images. This suggests that the GIST model may be using more local information. The reason for this could be that the GIST model uses local image features to create an overall global representation of the image. Since the GIST model uses line orientations to create a feature vector, it may be that blurred images degrade line orientations enough to make it difficult for the model to produce this global representation. Scrambled images, however, keep line orientations intact so that the GIST model can make these computations more accurately from what little information is available.

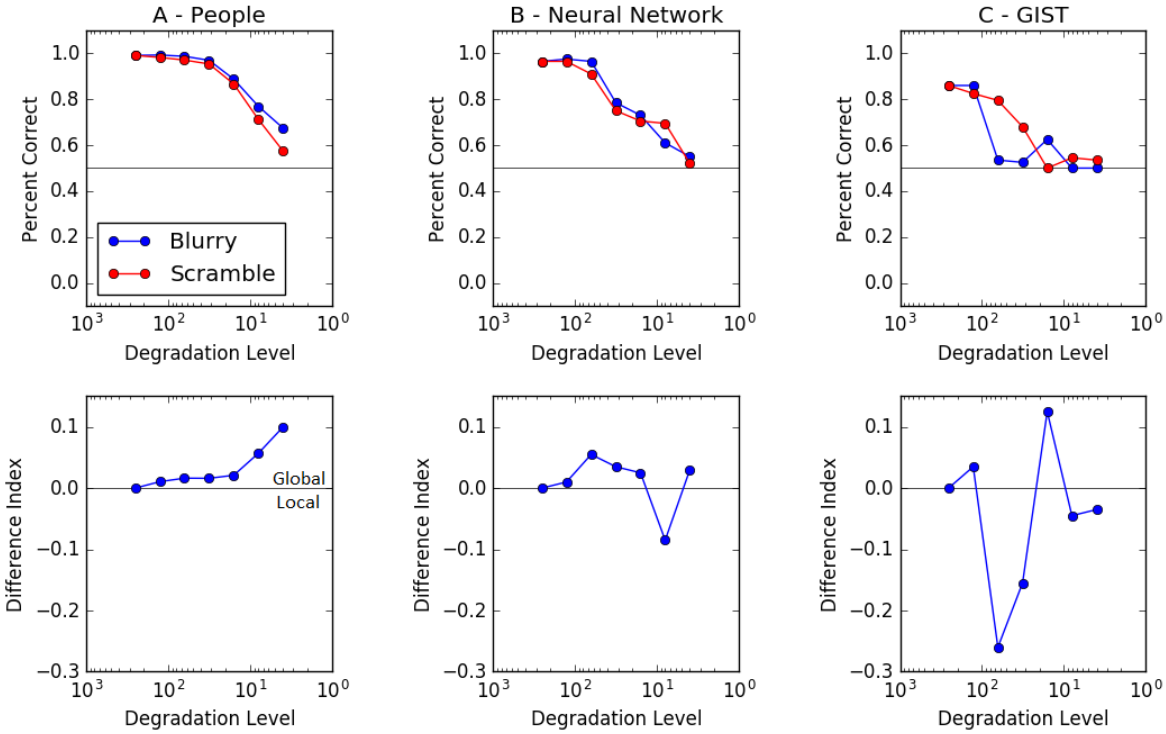


Figure 14: Local-Global Difference Index Measures for people, the neural network, and GIST model. The top three panels show classification acccuracies on blurred and scrambled images. The bottom three panels show the Difference Index for each of the perceptual systems.

10 Experiment Four - Training on Degraded Images

To make the neural network perform similarly to people, it may help to train on degraded images. To really mimic human behavior, the network must be as robust and invariant to image manipulations as people are. People have a much more vivid and robust visual experience than do neural networks. People can see visual scenes in all types of conditions that affect how distorted their visual field is. In this experiment, we train many neural networks, changing which degradations each network is trained on, and evaluate how this affects the neural network performance on the various degradation methods. The goal is to build the past visual experiene of the neural network so that it can be more robust to various degradations.

10.1 Methods

We trained each neural network by adding a custom layer in Caffe after the input layer. Each image in the batch currently being trained on is manipulated and the number of images in the batch is increased by a factor of how many degradation levels there are. For example, for a batch size of 32 images and using the scrambling distortion technique which has 7 levels of degradation, we increase the batch size by a factor of 8 to 256 images in the batch. This larger batch is then fed into the first convolutional layer of the network. During backpropagation, nothing needs to be done on this new input layer. For testing, this layer is removed and we tested on the same test set as has been used in previous sections of this paper.

10.2 Training on Scrambled Images

In this subsection, we present the results for training a 2-class neural network using the entire Places database, but adding gridded, scrambled with and without a grid as well as the intact images to the training set. The network was trained for 350,000 iterations. Figure 16 displays a comparison of the accuracies between the neural network trained on regular data and the neural network trained on the degraded images, as well as the results from the perceptual study.

As can be seen in Figure 15 this neural network becomes more reliant on local properties. Because the patches in a scrambled image still contain local object information, this is what the network can learn in order to classify these degraded images. Thus, global properties become less relevant and the neural network learns more local properties, resulting in a negative difference index.

10.3 Training on Noisy Images

Here, we trained a 2-class neural network using the entire Places database, but adding 4 noisy exemplars per each image. The network shown has been trained from scratch for 350,000 iterations. Figure 17 shows the results for a neural network trained on regular data, a neural network trained on regular and noisy data, and for people. The new neural network trained on degraded, noisy images performs overall much better at mid- and high-levels of degradation. This suggests that the neural network is more robust to noisy images. The original neural network was classifying these images around chance, choosing outdoor for all images. However, for noise levels of 0.2, the neural networks both achieve around the same classification accuracy.

As can be seen in Figure 15 this neural network has gotten more global, when looking at the Difference Index (which only takes into account blurred and scrambled images, not noisy images). Since adding noise to an image preserves global information, such as depth and openness in the scene, training on more noisy images would allow the neural network to learn these global features moreso than it learns local features, which are degraded. So, this result is expected even if the neural network was not explicitly trained on blurred images.

10.4 Training on Blurred Images

In this subsection, we trained the 2-class version of AlexNet on the Places database, including the blurred images from Experiment Three above (with the Butterworth filter). This neural network was trained from scratch for 170,000 iterations. Already, it is apparent that the network is classifying blurred images at a higher rate (see Figure 18). The network is getting closer to mirroring human performance. Also, the network is attuned to more global information, as it maintains a positive difference index (Figure 15). When images are blurred, objects and their properties are obscured by the blurring filter. Hence, the network has to learn overall features of the two image classes, which makes it more global in general.

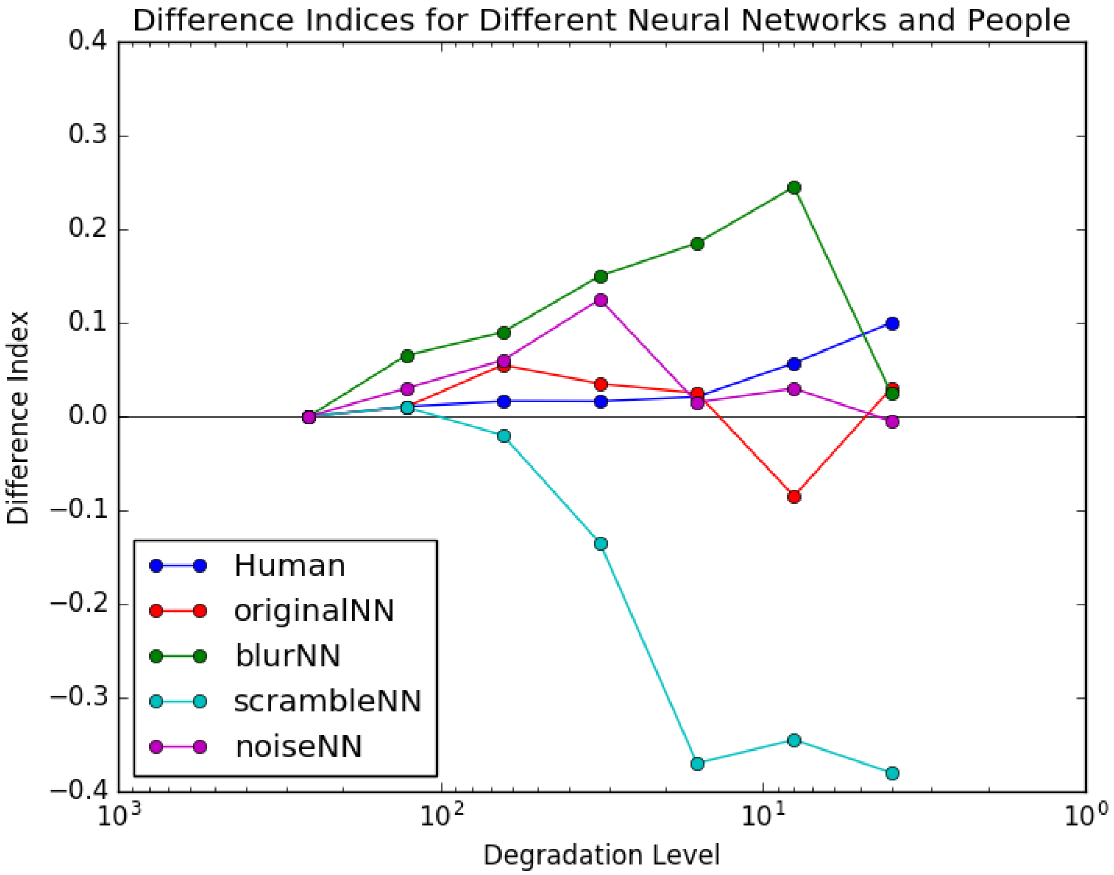

Difference Indices for Different Neural Networks and People
Difference Index
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
Degradation Level
10^3
10^2
10^1
10^0
Human
originalNN
blurNN
scrambleNN
noiseNN

Figure 15

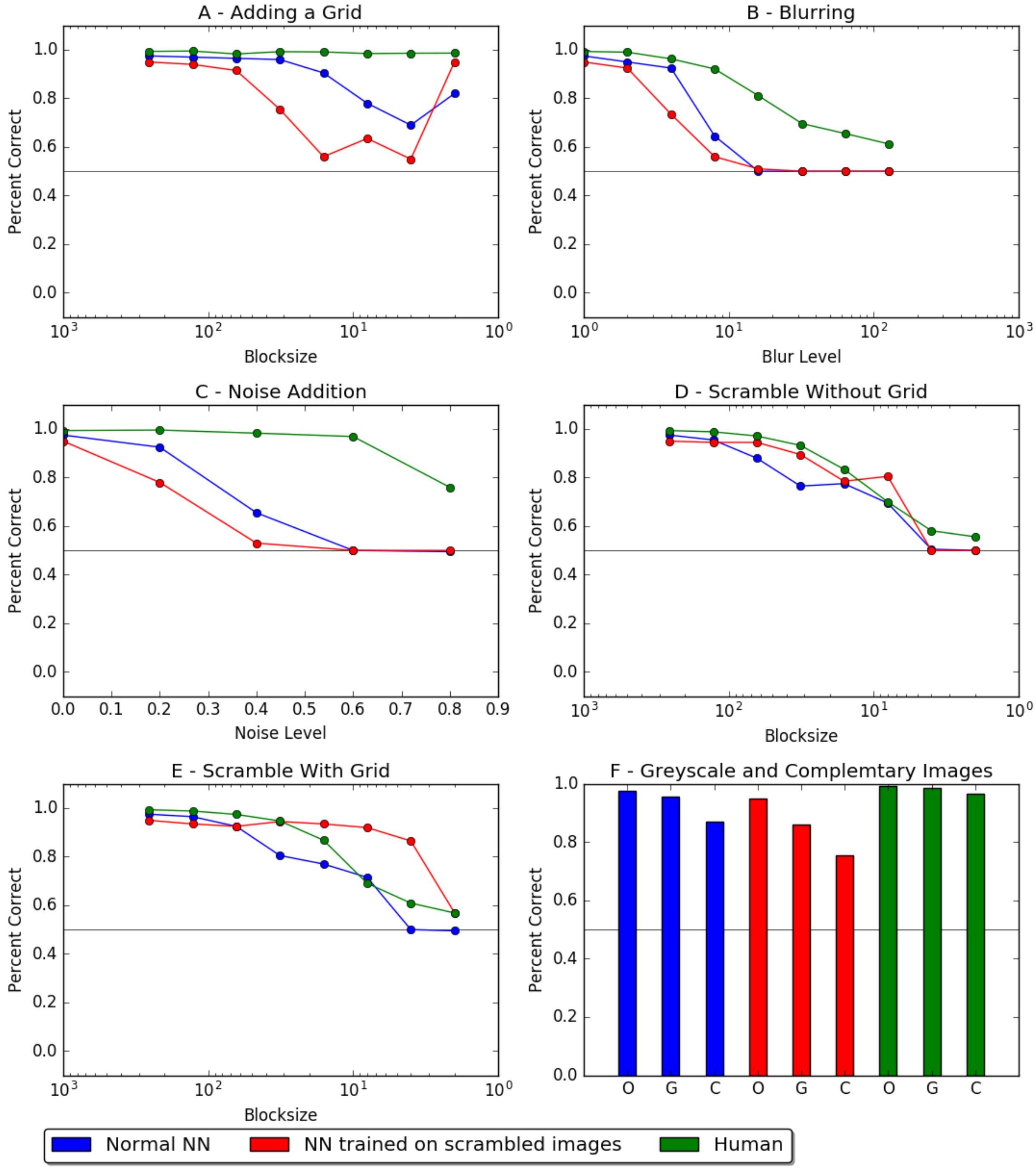


Figure 16: Percent accuracy for neural network trained on regular data, regular, scrambled (with and without a grid), and original images with a grid, and human results

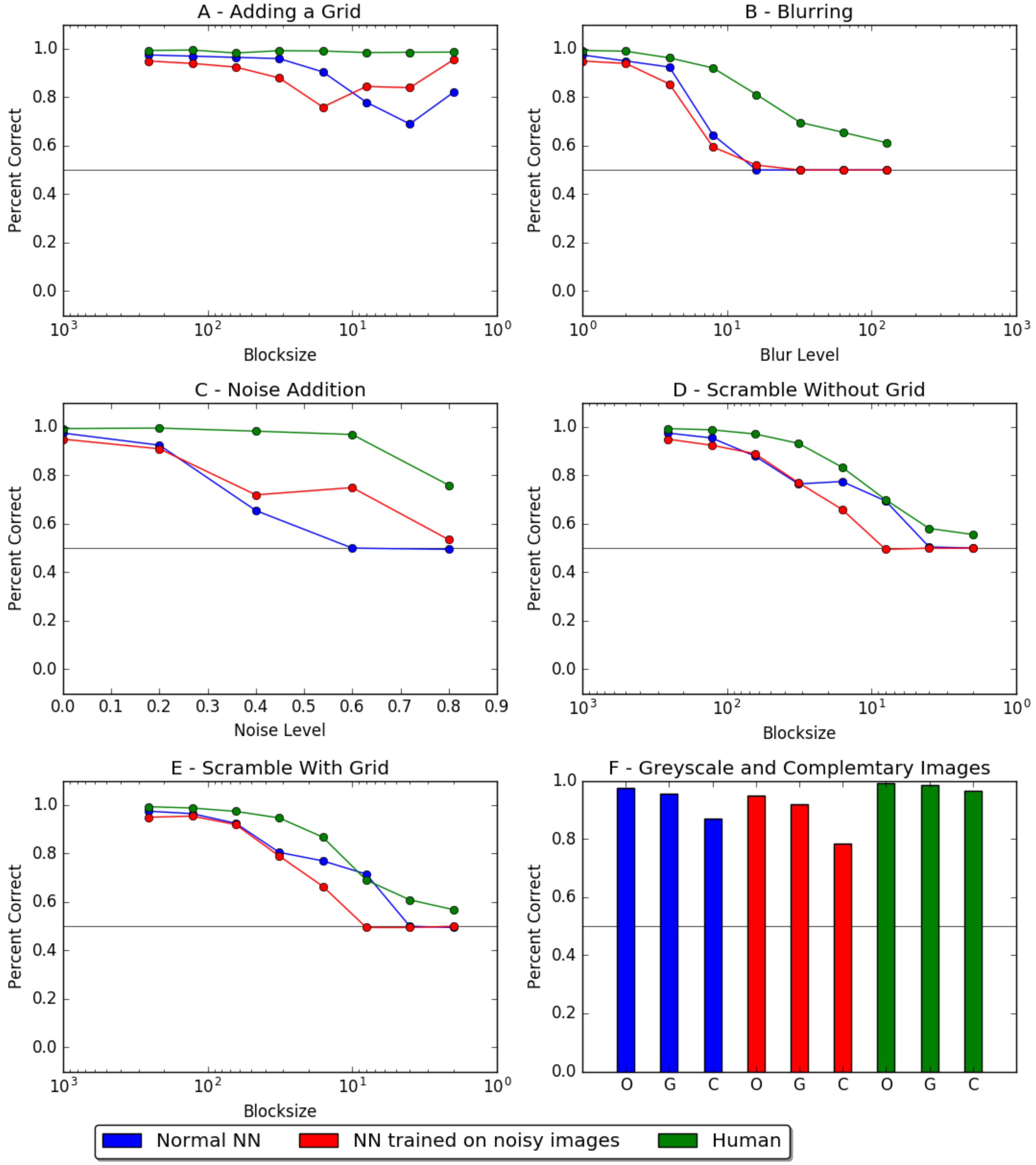


Figure 17: Percent accuracy for neural network trained on regular data, regular and noisy data, and human results

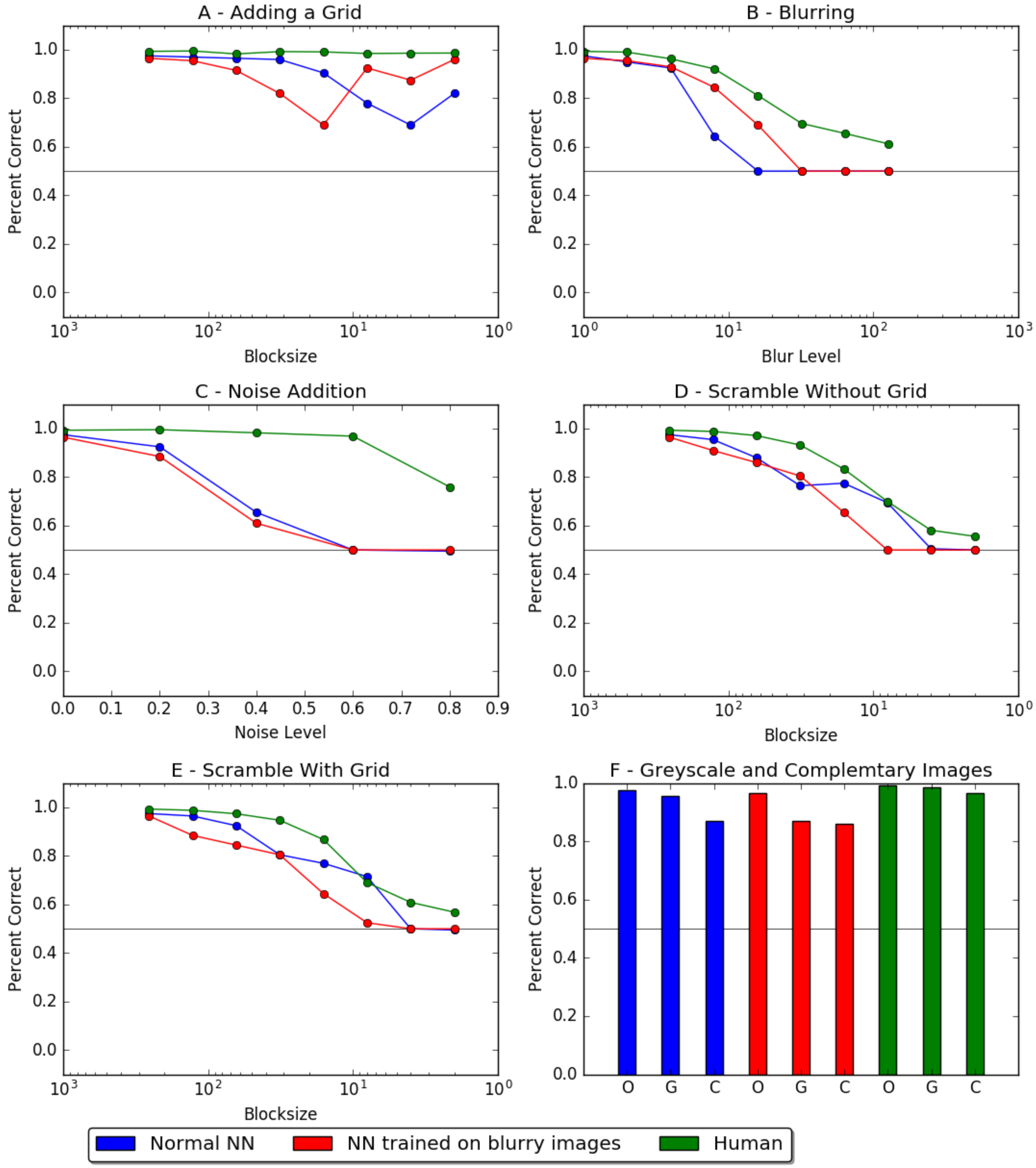


Figure 18: Percent accuracy for a neural network trained on regular data, a neural network trained on regular and blurry data, and human results

11 Discussion and Conclusion

In this paper, we looked at the role of local and global image information in classifying scenes as indoors or outdoors. It has been shown in previous work that both local and global image information are necessary for people to accurately interpret scene information, but that global information is perhaps collected and interpreted first. It has also been shown that classical machine learning techniques that respond to either local or global information can be used to mirror human performance on scene recognition tasks. In this paper, we compared a state-of-the-art neural network, whose design is inspired by the architecture of the human brain, to people's performance on an image classification task. We created an index and a mapping between locally and globally degraded images to determine if a given perceptual system is more local or global.

The results in Experiment One illustrate that both local and global information are important for people and machine vision to perform the indoor versus outdoor scene recognition task. For all types of perceptual systems, performance on the task falls as images become more and more degraded, no matter if those degradations methods reduce local or global information. People, however, are the most robust to degraded visual information, suggesting that they can recover lost or pertinent visual information from what is given to them. Neural networks and the GIST Descriptor model have more trouble recovering this information.

We also noticed that these computational algorthims have an extreme bias for selecting the outdoor image class, while people only have a slight bias for classifying images as outdoor. Intuitively, this makes sense as very blurred images have one overarching color and texture, similarly to outdoor images. Indoor images, on the contrary, posses many colors and textures as they contain many objects. This may indicate that object information may be relevant in helping machines and humans pick out indoor images but more global properties such as color and texture are more pertinent in picking out outdoor images.

In Experiment Two, we looked at ways to eliminate this bias in the case of neural networks. While the bias toward outdoor images was reduced by training on an equal number of indoor and outdoor images, in some cases the network became more biased toward indoor images. This does not replicate human performance, but it did make the neural network's average bias statistic closer to that of people.

To further compare neural networks to people, we created an index measure, in Experiment Three, called the Difference Index, in order to calculate how local or global a given perceptual system is. While people are better at classifying blurred images than they are with scrambled images, the neural network varies in how well it classifies each type of image. At higher degradation levels, the neural network is affected more by local information and does better at classifying scrambled images than blurred images. In order to build a computational replica of the human visual system, it is necessary to create a neural network that more closely mirrors the Difference Index for human beings, so that performance on degraded images is similar. One area to look at for improving the neural network is the blurring filter. While most of our Butterworth filters attenuated half the amplitude at the frequency cutoffs, the two smallest cutoffs were at a little less than half. It may be interesting to test different filter orders and different types of filters to try and get a closer matching between blurred and scrambled images.

In Experiment Four, we looked to see if neural networks can be trained to be able to recover the lost visual information. In the case of training on scrambled images, the new neural network surpassed human performance on more degraded blocksizes. It is unclear if these neural networks trained on manipulated images are learning new features that may correspond to local or global image features, or if they are just being tuned to where information may lie in the image. For example, scrambled images are jumbled up at known points depending on the blocksize. The neural network can ignore the relation of these blocks to one another and just capture information in each block, possibly having representations for capturing information in blocks of different sizes. If this is true, then it would still be the case that the neural network is learning local image features.

In the case of blurred and noisy images, which are global manipulations, the neural network does not need to compute filters for different areas in the image since the relationship among patches in the image will remain, but instead learns new filters to capture different features. These features may correspond to more global features which could explain how it performs better at classifying these globally distorted images than the normal neural network.

Further studies that can be conducted to try and build a neural network model that resembles the human visual system include changing the architecture of the neural network. It is possible that neural networks can be built to possess either a more globally-respondent or more locally-respondent architecture, based on the size of the filters in the convolutional layer and the number of convolutional layers and fully connected layers. Changing the AlexNet architecture to resemble a more global or more local architecture may lead to interesting results that further corroborates or rejects the similarity of neural networks to the human brain.

In general, there are many interesting questions in the realm of human perception and scene recognition. While neural networks are approaching human performance, they are still not as robust. It seems that human beings can take a scene with missing information and extract this missing information from their experience and past knowledge of how visual scenes are supposed to look. Neural networks, however, perform scene classifcation by learning features in a large set of images that help them perform this task. One area to look to solve this problem may be recurrent neural networks for scene recognition, which do have a time-dependent representation.

For human beings, it is still unclear how top-down encoding of past experiences affects visual processing of scenes with missing information. A closer look at the brain while people process degraded images may reveal how this process works and what exactly is needed to make up a scene with degraded information.

12 Acknowledgements

This research was funded through the generosity of the Kaminsky Family Fund Award. I am also grateful for the time and effort put into providing comments and feedback as well as pointers for my thesis research by Professor Emily Cooper. I also appreciate the edits and support from the people in my thesis group - Irene Feng and Jonathan Huang.

References

[1] Moshe Bar and Shimon Ullman, *Spatial context in recognition*, Perception **25** (1996), no. 3, 343–352.

[2] Irving Biederman, *On the semantics of a glance at a scene*, 1981.

[3] Irving Biederman, *Perceiving real-world scenes*, Science **177** (1972), no. 4043, 77–80.

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

[5] Susan J Boyce, Alexander Pollatsek, and Keith Rayner, *Effect of background information on object identification*, Journal of Experimental Psychology: Human Perception and Performance **15** (1989), no. 3, 556–566.

[6] Monica S Castelhano and Chelsea Heaven, *Scene context influences without scene gist: Eye movements guided by spatial associations in visual search*, Psychonomic bulletin & review **18** (2011), no. 5, 890–896.

[7] Radoslaw M Cichy, Aditya Khosla, Dimitrios Pantazis, Antonio Torralba, and Aude Oliva, *Deep neural networks predict hierarchical spatio-temporal cortical dynamics of human visual object recognition*, arXiv preprint arXiv:1601.02970 (2016).

[8] Li Fei-Fei, Asha Iyer, Christof Koch, and Pietro Perona, *What do we perceive in a glance of a real-world scene?*, Journal of vision **7** (2007), no. 1, 1–29.

[9] Mojgan M. Gorkani and Rosalind W. Picard, *Texture orientation for sorting photos "at a glance"*, Pattern Recognition, 1994. Vol. 1-Conference A: Computer Vision & Image Processing., Proceedings of the 12th IAPR International Conference on, vol. 1, IEEE, 1994, pp. 459–464.

[10] Michelle R Greene and Aude Oliva, *Recognition of natural scenes from global properties: Seeing the forest without representing the trees*, Cognitive psychology **58** (2009), no. 2, 137–176.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, *Delving deep into rectifiers: Surpassing human-level performance on imagenet classification*, Proceedings of the IEEE international conference on computer vision, 2015, pp. 1026–1034.

[12] John M Henderson and Andrew Hollingworth, *High-level scene perception*, Annual review of psychology **50** (1999), no. 1, 243–271.

[13] John M Henderson and Andrew Hollingworth, *Eye movements during scene viewing: An overview*, Eye guidance in reading and scene perception **11** (1998), 269–293.

[14] Jeffrey S Johnson and Bruno A Olshausen, *Timecourse of neural signatures of object recognition*, Journal of Vision **3** (2003), no. 7, 4–4.

[15] Nikolaus Kriegeskorte, *Deep neural networks: a new framework for modeling biological vision and brain information processing*, Annual Review of Vision Science **1** (2015), 417–446.

[16] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce, *Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories*, Computer vision and pattern recognition, 2006 IEEE computer society conference on, vol. 2, IEEE, 2006, pp. 2169–2178.

[17] David Marr, *A computational investigation into the human representation and processing of visual information*, WH San Francisco: Freeman and Company **1** (1982), no. 2.

[18] David Navon, *Forest before trees: The precedence of global features in visual perception*, Cognitive psychology **9** (1977), no. 3, 353–383.

[19] Aude Oliva, *Gist of the scene*, Neurobiology of attention **696** (2005), no. 64, 251–258.

[20] Aude Oliva and Philippe G Schyns, *Diagnostic colors mediate scene recognition*, Cognitive psychology **41** (2000), no. 2, 176–210.

[21] Aude Oliva and Phillipe G Schyns, *Coarse blobs or fine edges? evidence that information diagnosticity changes the perception of complex visual stimuli*, Cognitive psychology **34** (1997), 72–107.

[22] Aude Oliva and Antonio Torralba, *Modeling the shape of the scene: A holistic representation of the spatial envelope*, International journal of computer vision **42** (2001), no. 3, 145–175.

[23] Aude Oliva and Antonio Torralba, *Building the gist of a scene: The role of global image features in recognition*, Progress in brain research **155** (2006), 23–36.

[24] Devi Parikh, *Recognizing jumbled images: The role of local and global information in image classification*, Computer Vision (ICCV), 2011 IEEE International Conference on, IEEE, 2011, pp. 519–526.

[25] Mary C Potter, *Meaning in visual search*, Science **187** (1975), no. 4180, 965–966.

[26] Mary C Potter, Adrian Staub, Janina Rado, and Daniel H O'connor, *Recognition memory for briefly presented pictures: the time course of rapid forgetting*, Journal of Experimental Psychology: Human Perception and Performance **28** (2002), no. 5, 1163–1175.

[27] Ariadna Quattoni and Antonio Torralba, *Recognizing indoor scenes*, Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, IEEE, 2009, pp. 413–420.

[28] Laura W. Renninger and Jitendra Malik, *When is scene identification just texture recognition?*, Vision research **44** (2004), no. 19, 2301–2311.

[29] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K Reiter, *Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition*, Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, ACM, 2016, pp. 1528–1540.

[30] Yi Sun, Yuheng Chen, Xiaogang Wang, and Xiaoou Tang, *Deep learning face representation by joint identification-verification*, Advances in neural information processing systems, 2014, pp. 1988–1996.

[31] Martin Szummer and Rosalind W Picard, *Indoor-outdoor image classification*, Content-Based Access of Image and Video Database, 1998. Proceedings., 1998 IEEE International Workshop on, IEEE, 1998, pp. 42–51.

[32] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf, *Deepface: Closing the gap to human-level performance in face verification*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1701–1708.

[33] Antonio Torralba, Kevin P Murphy, and William T Freeman, *Contextual models for object detection using boosted random fields*, Neural Information Processing Systems, vol. 1, 2004, pp. 1401–1408.

[34] Barbara Tversky and Kathleen Hemenway, *Categories of environmental scenes*, Cognitive psychology **15** (1983), no. 1, 121–149.

[35] Julia Vogel and Bernt Schiele, *A semantic typicality measure for natural scene categorization*, Joint Pattern Recognition Symposium, Springer, 2004, pp. 195–203.

[36] Julia Vogel, Adrian Schwaninger, Christian Wallraven, and Heinrich H Bülthoff, *Categorization of natural scenes: Local versus global information and the role of color*, ACM Transactions on Applied Perception (TAP) **4** (2007), no. 3, 19.

[37] Li Wan, Matthew Zeiler, Sixin Zhang, Yann L Cun, and Rob Fergus, *Regularization of neural networks using dropconnect*, Proceedings of the 30th International Conference on Machine Learning (ICML-13), 2013, pp. 1058–1066.

[38] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba, *Sun database: Large-scale scene recognition from abbey to zoo*, Computer vision and pattern recognition (CVPR), 2010 IEEE conference on, IEEE, 2010, pp. 3485–3492.

[39] Daniel L Yamins, Ha Hong, Charles Cadieu, and James J DiCarlo, *Hierarchical modular optimization of convolutional networks achieves representations similar to macaque it and human ventral stream*, Advances in neural information processing systems, 2013, pp. 3093–3101.

[40] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva, *Learning deep features for scene recognition using places database*, Advances in neural information processing systems, 2014, pp. 487–495.

13 Appendix

Figures 19 - 21 demonstrate the results from running Experiment One on the GIST LDA model rather than the GIST SVM model. Figures 22 - 24 show the plots for comparing results from the perceptual study, the pre-trained Places205-AlexNet, and our neural network trained for the indoor/outdoor classification task. Table 5 compares the pre-trained AlexNet with our results from training AlexNet on 205 scene classes in Places205.

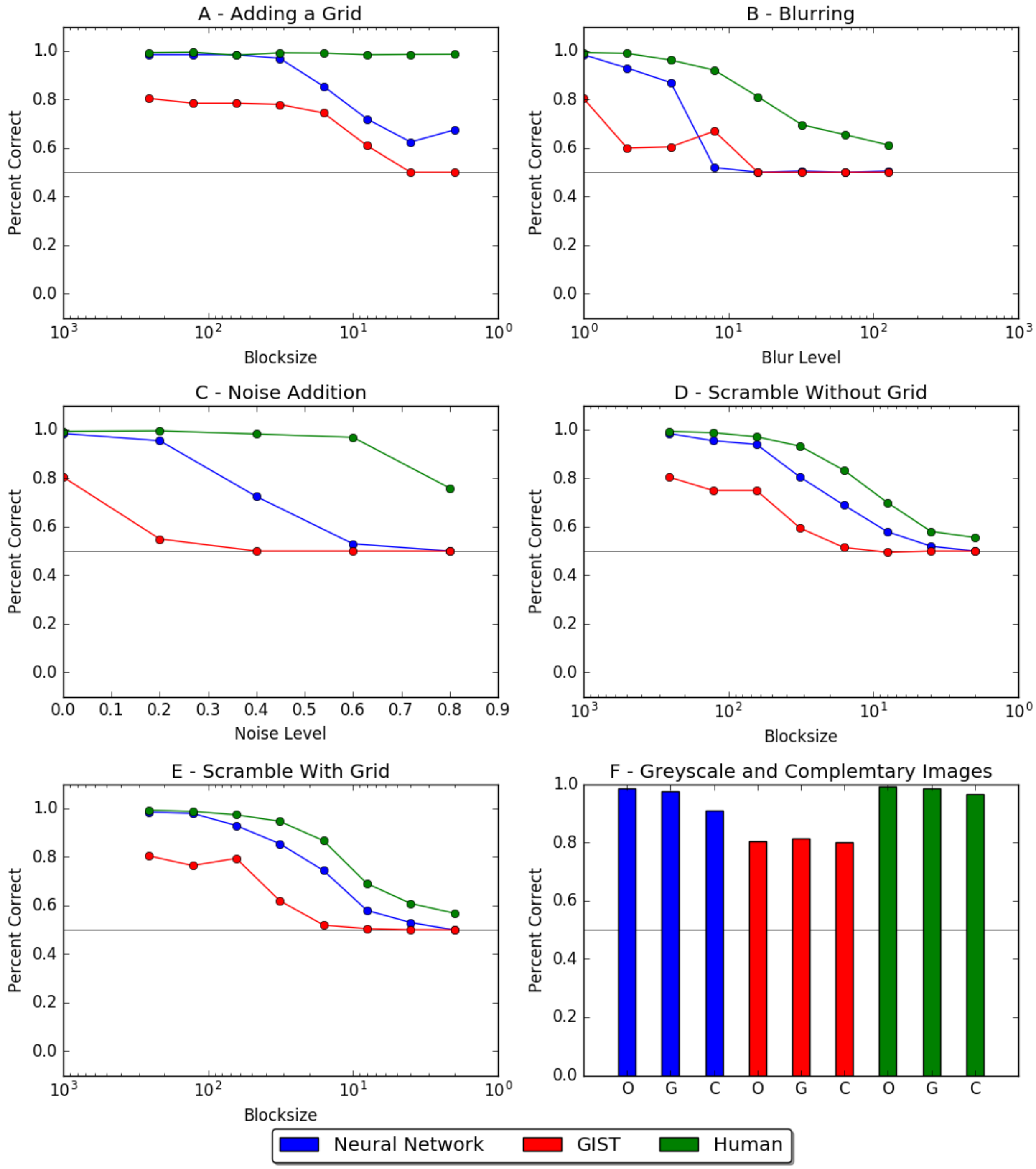


Figure 19: Percent correct by image manipulation for several perceptual systems - a neural network, a GIST LDA model, and human subjects. The x-axis is oriented such that low-level/no manipulations are on the left and higher-level manipulations are on the right. In Panel F, O stands for original images, G for greyscale images, and C for complementary images. The black horizontal line in each plot represents chance.

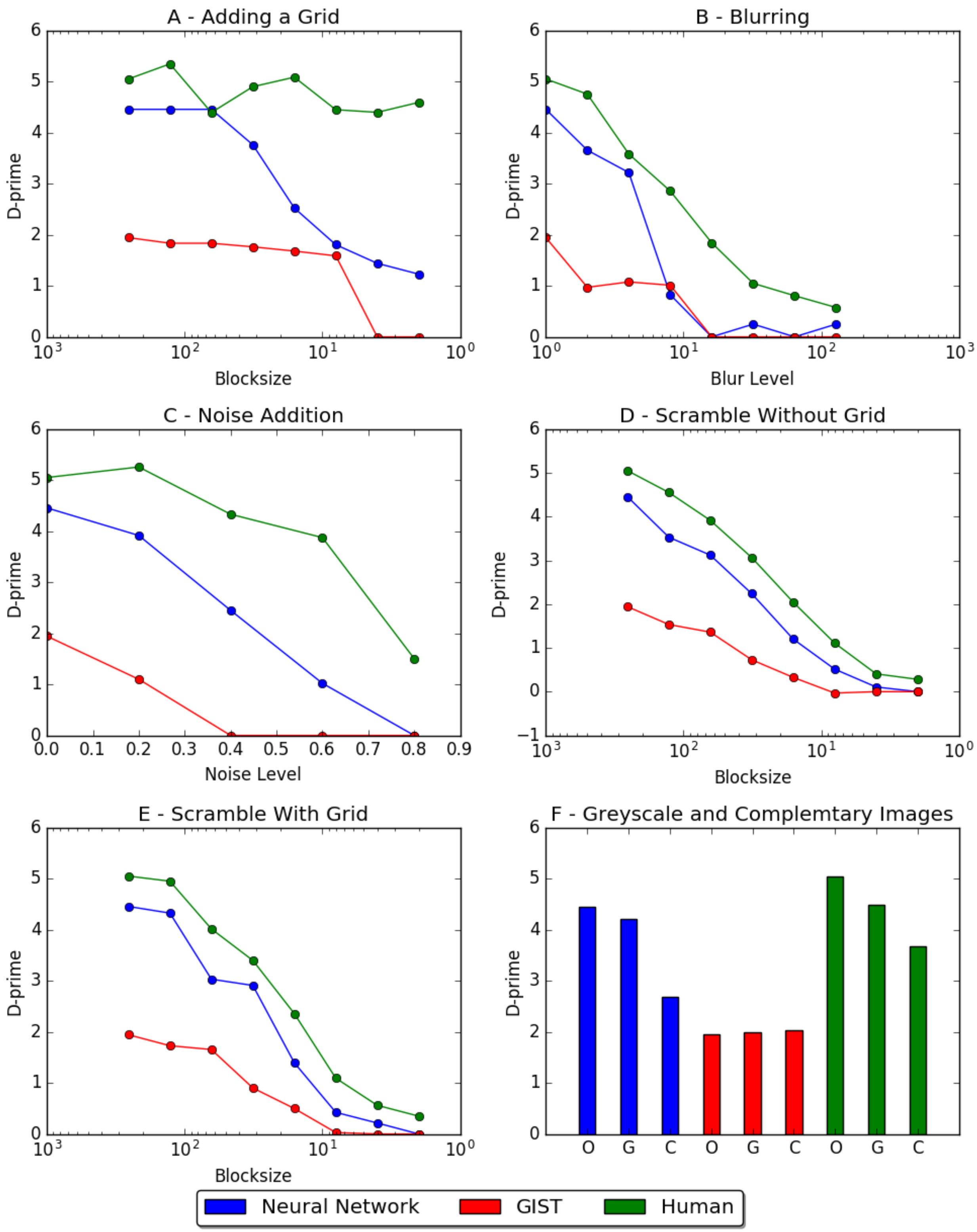


Figure 20: D-prime statistics for the neural network model, GIST LDA model, and human subjects on various image manipulations. Values closer to 0 indicate a strong bias, as a d-prime value of 0 implies that the hit rate is equal to the false alarm rate. In Panel F, O stands for original images, G for greyscale images, and C for complementary images.

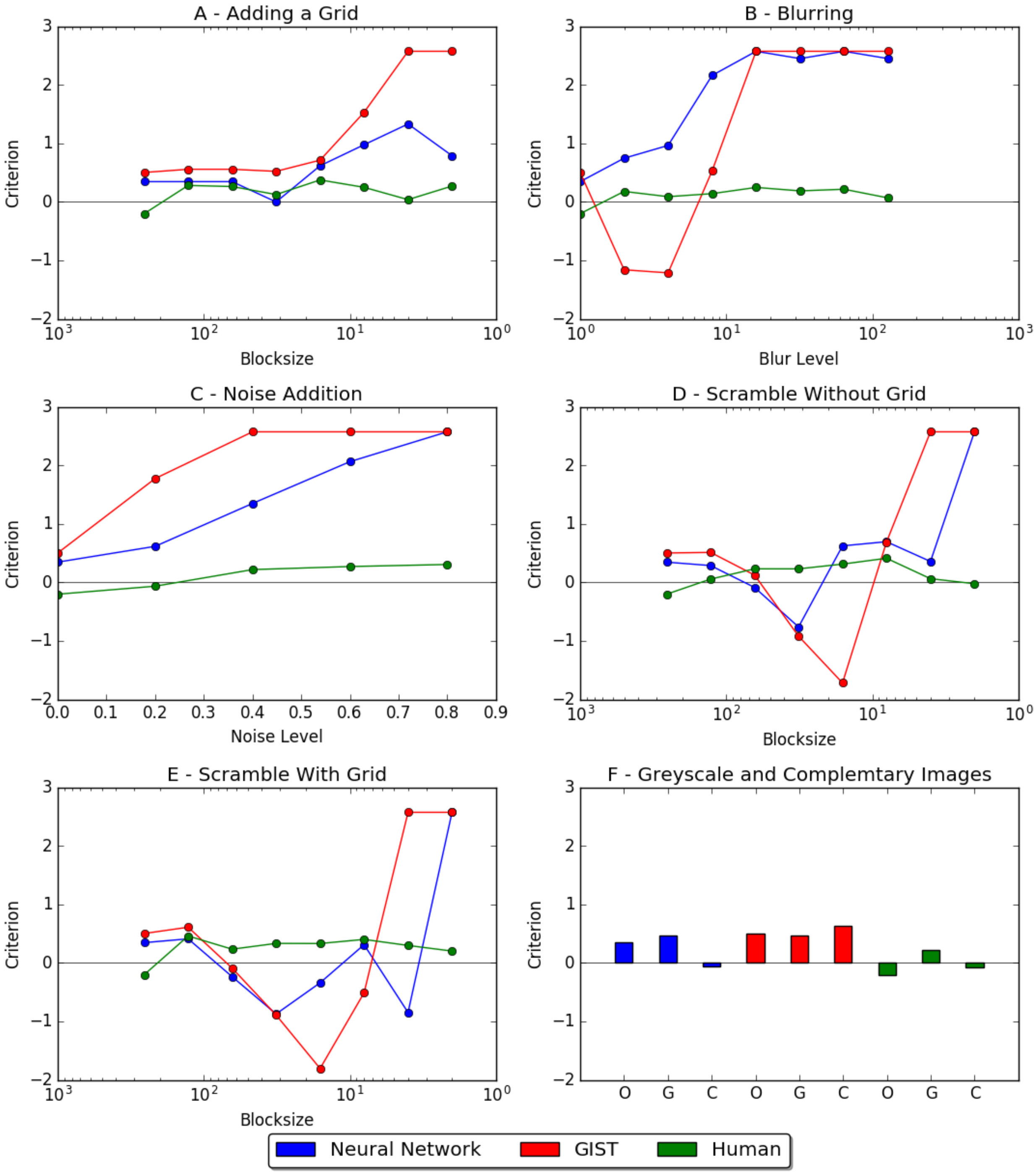


Figure 21: Criterion statistics for the neural network, GIST LDA model, and human subjects on various image manipulations. High positive values indicate an outdoor bias whereas high negative values indicate an indoor bias. The black horizontal line indicates no bias (value 0).

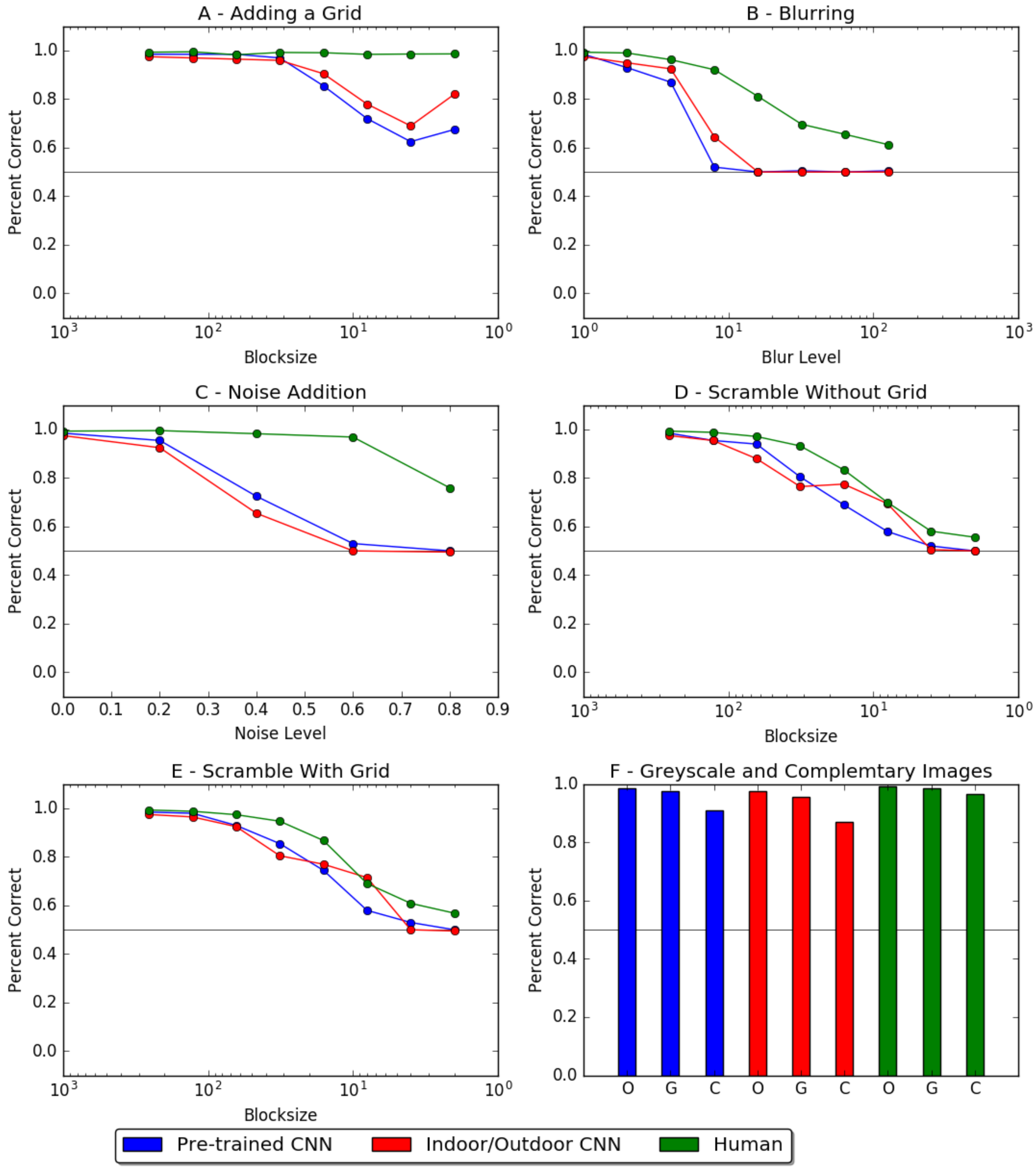


Figure 22: Percent accuracy by image distortion of several perceptual systems - a pre-trained neural network for 205 scene classes, a neural network trained specifically for the indoor/outdoor classification task, and human subjects

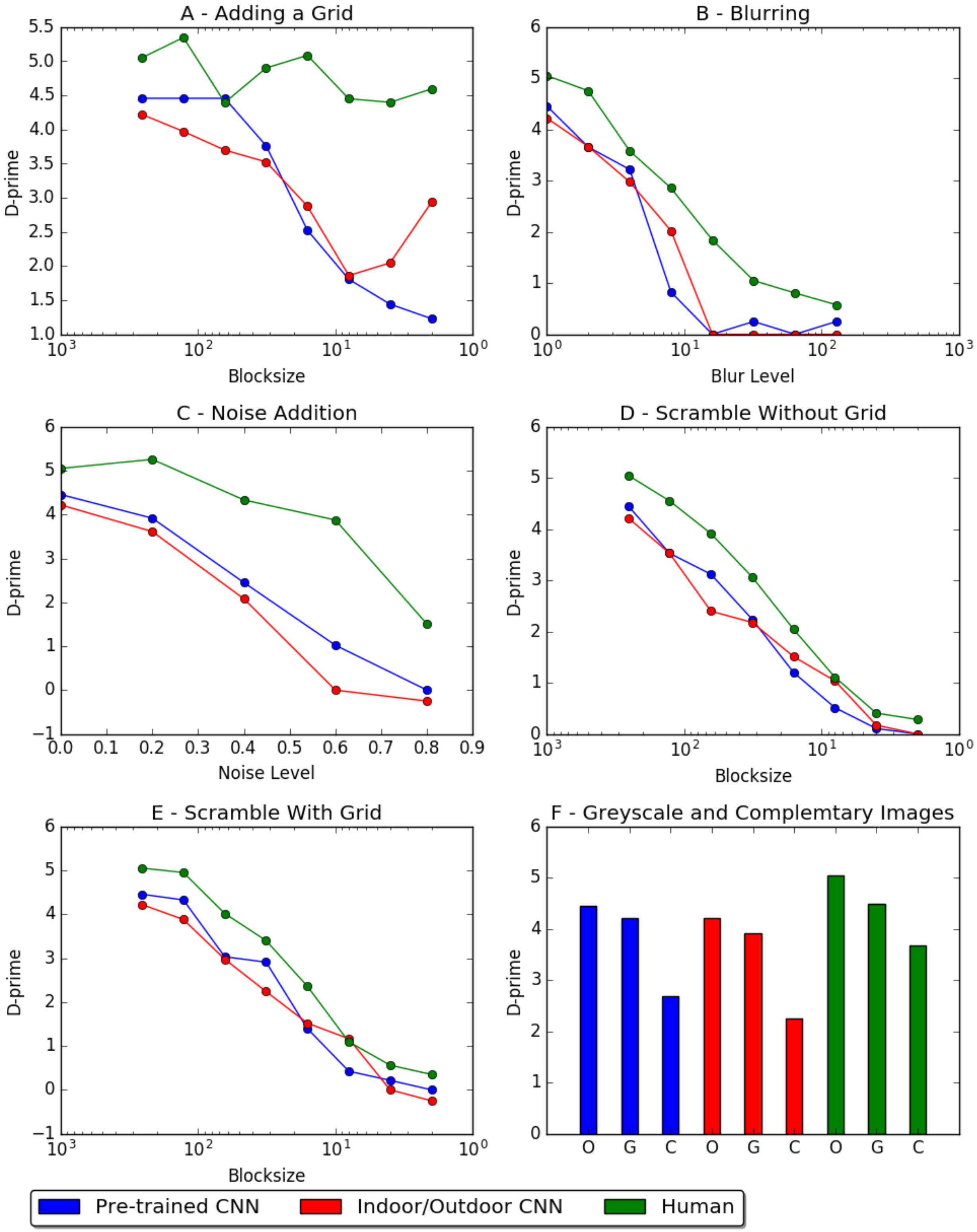


Figure 23: D-prime statistics of the pre-trained neural network, indoor/outdoor neural network, and human subjects on various distortion methods. Values closer to 0 indicate a strong bias, as a d-prime value of 0 implies that the hit rate is equal to the false alarm rate.

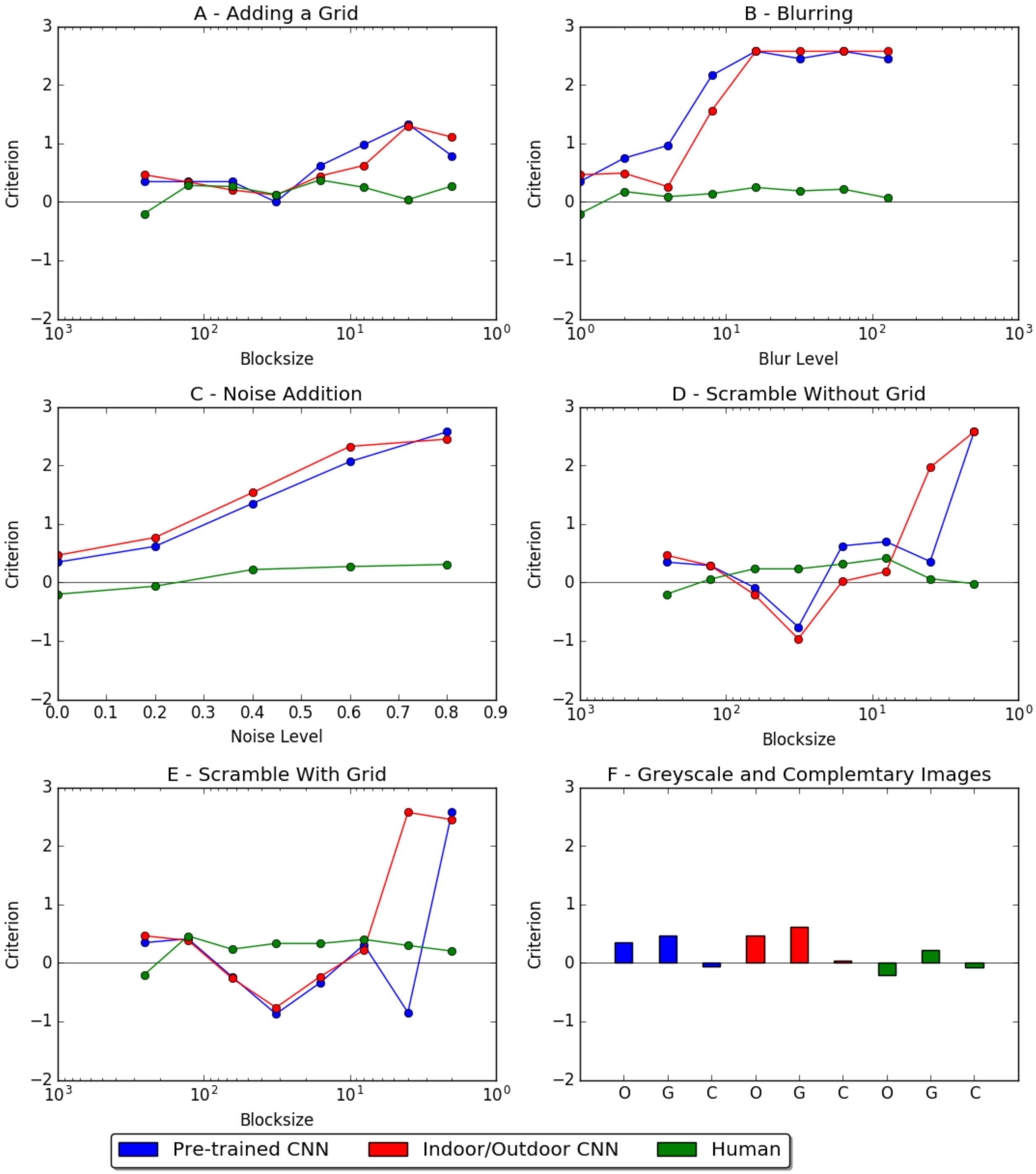


Figure 24: Criterion statistics of the pre-trained neural nentwork, the indoor/outdoor neural network, and human subjects on various distortion methods. High positive values indicate an outdoor bias whereas high negative values indicate an indoor bias.

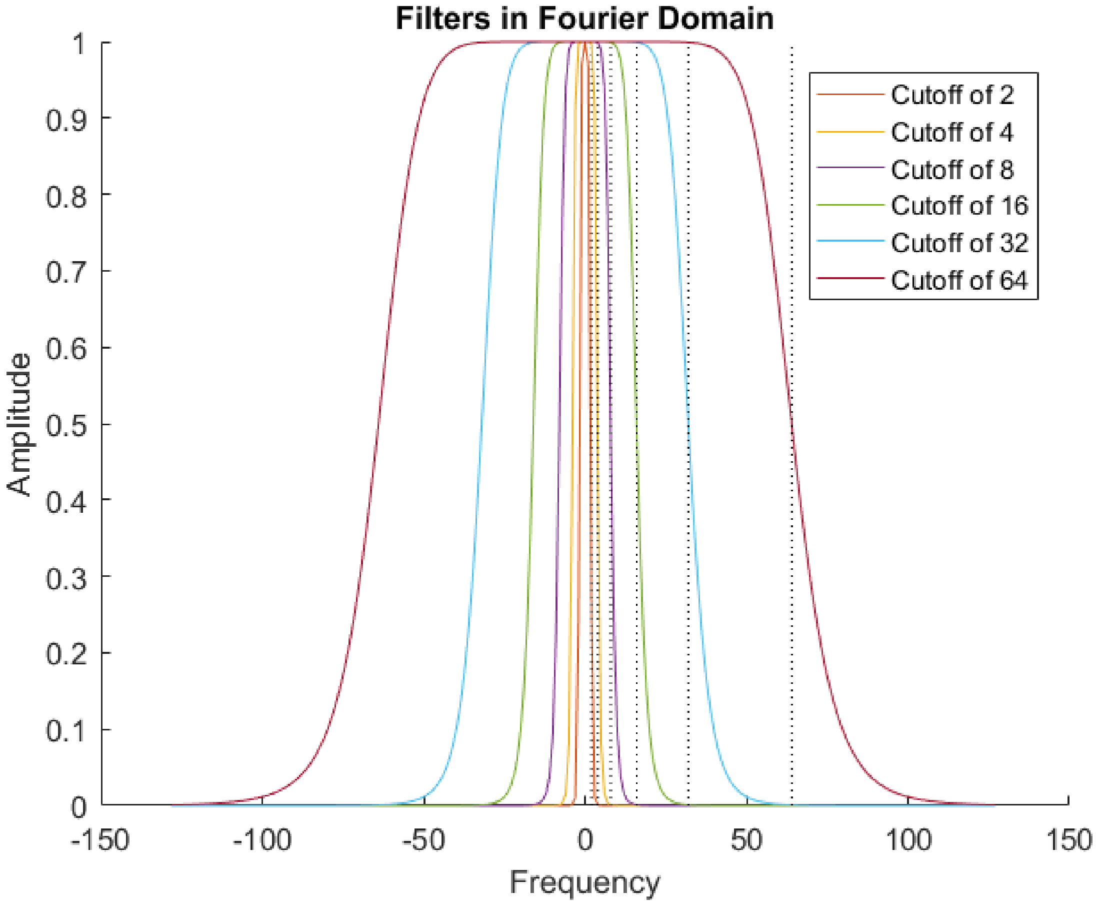


Figure 25: Butterworth Filter Behaviors for different frequency cutoffs

Table 5: Pre-trained Neural Network vs. Our Trained Neural Network (both with 205 scene classes). Each value is a percentage of images answered correctly.

	Pre-trained NN	Our NN		Pre-trained NN	Our NN
Original	98.5	99	Blurred-2	93	96.5
Grid-2	67.5	75	Blurred-4	87	81.5
Grid-4	62.5	83.5	Blurred-8	52	56.5
Grid-8	72	89.5	Blurred-16	50	50.5
Grid-16	85.5	90.5	Blurred-32	50.5	50
Grid-32	97	88	Blurred-64	50	50.5
Grid-64	98.5	97.5	Blurred-128	50.5	50
Grid-128	98.5	99	Noisy-0.2	95.5	95
Scrambled-2	50	52	Noisy-0.4	72.5	70
Scrambled-4	53	59	Noisy-0.6	53	51.5
Scrambled-8	58	60	Noisy-0.8	50	50
Scrambled-16	74.5	68	Grayscale	97.5	96.5
Scrambled-32	85.5	76	Complement	91	84.5
Scrambled-64	93	91.5			
Scrambled-128	98	98			