SATBench: A Benchmark of the Human Speed-Accuracy Tradeoff in Recognizing Objects

Ajay Subramanian
New York University
as15003@nyu.edu

Omkar Kumbhar
New York University
omkar.kumbhar@nyu.edu

Elena Sizikova
New York University
es5223@nyu.edu

Najib J. Majaj
New York University
najib.majaj@nyu.edu

Denis G. Pelli
New York University
denis.pelli@nyu.edu

Abstract

People take a variable amount of time, 0.1 to 10 s, to recognize an object. The reaction time depends on the stimulus and task, and people can trade off speed for accuracy. That tradeoff is a crucial human skill. Neural networks exhibit high accuracy in object recognition, but most current models cannot dynamically adapt to respond with less computation, which is a problem in time-sensitive applications like driving. Towards the goal of using networks to model how people recognize objects, we here present a benchmark dataset (with model fits) of the human speed-accuracy tradeoff (SAT) in recognizing CIFAR-10 [1] and STL-10 [2] images. In each trial, a beep, indicating the desired reaction time, sounds at a fixed delay after the target onset, and the observer's response counts only if it occurs near the time of the beep. With practice, observers quickly learn to respond at the time of the beep. In a series of blocks, we test many beep latencies, i.e., reaction times. We observe that human accuracy increases with reaction time, and we compare its characteristics with the behavior of several dynamic neural networks that can trade off speed and accuracy. After limiting the network resources and adding image perturbations (grayscale conversion, noise, blur) to bring the two observers (human and network) into the same accuracy range, we show that humans and networks exhibit very similar tradeoffs. We conclude that dynamic neural networks are a promising model of human reaction time in recognition tasks. Our dataset[1] and code[2] are publicly available.

1 Introduction

Unlike neural networks, a typical and salient feature of human behavior is the ability to flexibly tradeoff accuracy for speed, which is called the speed-accuracy tradeoff (SAT). Here, we argue that SAT is crucial, both for deployment of machine learning in time-sensitive applications, and for better understanding of human decision making. We present a benchmark of SAT in human object recognition and propose to model its properties with dynamic neural networks. Our dataset includes analysis of fits by three neural networks.

As signal strength (e.g., contrast) increases, humans respond more quickly and more accurately, and there is a tight relation between signal sensitivities measured by accuracy or by reaction time. Palmer et al. [3] showed that a diffusion model of perceptual decision making could account for the relation.

[1] See https://osf.io/zkvep/ for dataset.

[2] See https://github.com/ajaysub110/anytime-prediction for code.

Submitted to the 35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks. Do not distribute.

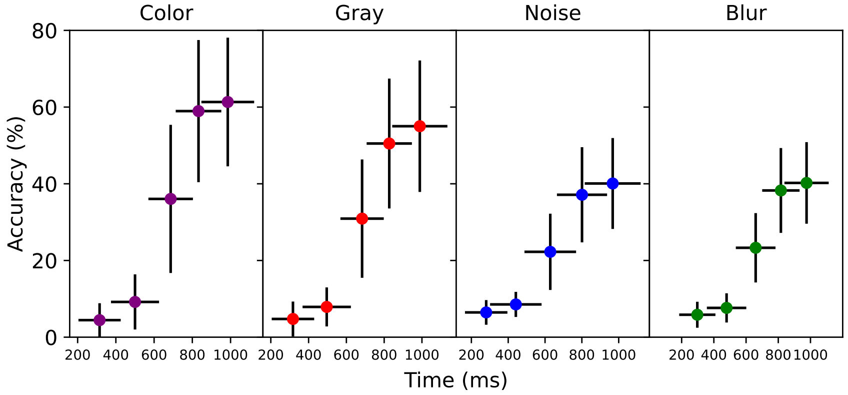


Figure 1: *Scatter plots showing mean and standard deviation of accuracy and reaction time across participants for several image perturbations*. For each experiment, each of the five points corresponds to a block of trials that required the participant to respond within a small range oft duration centered on: 200 ms, 400 ms, 600 ms, 800 ms, or 1000 ms. With more time, human observers classify more accurately. That is the speed-accuracy tradeoff (SAT). The title of each graph (Color, Gray, Noise, Blur) refers to an image degradation that is explained below. Corresponding results on the STL-10 [2] dataset are shown in Figure 9.

In Figure 1, we show typical speed-accuracy tradeoffs observed in the human data we have collected, when subjects are presented with color, grayscale, noisy, and blurry images. Accuracy decreases gradually as allowed reaction time is reduced, which allows observers to make reasonable decisions even with limited time. Neural networks are currently extremely popular computational models due to their excellent accuracy in tasks such as pattern recognition [4], medical data analysis [5], robotics [6], and many others [7]. However, most recent neural network models are trained to use a fixed number of layers to make decisions and therefore cannot adapt to unexpected time constraints [8]. Due to increased use of networks in wearable sensor technologies for health monitoring [9] or in applications such as obstacle and pedestrian avoidance in autonomous driving [10], it is crucial that these computational models offer reasonable accuracy even when time available for inference is reduced. In order to teach models to “fail gracefully", i.e., offer partial accuracy with partial time (or FLOPS), we look to human SAT as a successful example of this ability. Taking human performance as a goal for machine development was key in the development of neural networks [11, 12], and continues to motivate developments in artificial intelligence (AI) research [13, 14, 15, 16]. On the other hand, finding an accurate model of human decision making under time pressure would be a milestone in neuroscience and might be a first step toward understanding slow reading, which is primarily characterized by very slow performance.

Inspired by these works [17, 15, 14], we benchmarked the SAT of humans recognizing objects to provide a dataset for modeling of this human ability. Within each block of trials, the observer is taught to respond at a different fixed latency. Each block yields a point in a plot of accuracy vs. reaction time, and the responses from many blocks trace out the speed-accuracy tradeoff. The task is to identify the predefined category (1 of 10) of an image from the CIFAR-10 [1] collection of natural images, which are commonly used to benchmark computer vision algorithms. As models of the human tradeoff, we have evaluated three recent computational networks that allow early exits and adaptive computation as ways to vary computational effort. The first model is a convolutional recurrent neural network (ConvRNN), introduced by [18] which has already been used to model the human speed-accuracy tradeoff. This model relies on confidence saturation as an exit strategy to dynamically throttle computation. The other two models, MSDNet [19] and SCAN [20], are both popular dynamic-depth, anytime-prediction models that are used for computer vision and related applications. For human and network, We measured accuracy and time (or FLOPS) in classifying degraded CIFAR-10 images. To compare the speed-accuracy tradeoffs of networks and humans, we assume a linear correspondence between reaction time in milliseconds (ms) and the number of floating point operations (FLOPS) consumed by the network. The offset and slope of the linear correspondence are determined by linear regression. We correlate networks against human accuracy across reaction time. Our results indicate that anytime prediction is a promising model for human accuracy and reaction time in object recognition because it achieves a high correlation with the human tradeoff. Our contributions are:

- We study how human observers recognize objects, i.e., identify the class (1 of 10) of each image in the CIFAR-10 [1] and STL-10 [2] datasets, under less-than-ideal viewing conditions. Our main contribution is an open-access dataset for the human speed-accuracy tradeoff (SAT) in object recognition. This dataset, gathered from psychophysical experiments with 142 subjects, spans a wide range of classification accuracies (20% to 90%) under several image perturbations: grayscale conversion, blur, and noise. It is intended for comparison with computational models of visual recognition.

- We evaluate the ability of several artificial neural networks to capture the characteristics of human SAT, and show that the MSDNet [21] dynamic-depth neural network matches human SAT better than previous work [18].

- We perform an extensive quantitative comparison between speed-accuracy tradeoffs in humans and several artificial neural networks. In doing so, we introduce two metrics, an accuracy-range metric and a correlation metric, which ease comparison of model and human performance.

2 Related work

Comparing humans and neural networks. Human vision inspired early neural networks [11, 12] that incorporate some computational features of human vision [22]. Many properties of neural networks, such as filters [23] and attention [24], were inspired by the human brain. Recent studies [14] suggest more properties that neural networks might learn from humans, and in this work, we focus on SAT. We look at the class of networks that can vary their computational effort, and thus model human SAT. In machine learning literature, these models are known as dynamic neural networks [8]. They adapt their architecture to the challenge of input data to reduce the mean cost of inference [19, 25, 26, 27, 28, 21, 29, 30]. Many applications of networks, such as analysis during autonomous driving [10] and mobile health sensors [9] are time-sensitive, and require reasonable accuracy even with brief time and few FLOPS. Taking humans as a good example of speed-accuracy tradeoff, we here record a benchmark. We assess several models of the SAT, including two recent dynamic depth networks [21, 20] and a recurrent network [18]. We hope releasing a SAT benchmark will encourage future experimentation with different models [31].

Measuring the speed-accuracy tradeoff (SAT). Given more time, people generally do better. McElree and Carrasco [32] analyzed the speed-accuracy tradeoff in humans on a visual search task, in which observers tried to find a target in an array of distractors. They manipulated task difficulty by adding more distractors. Figure 1 shows human object recognition accuracy on CIFAR-10 images as a function of reaction time [1]. Mirzaei et al. [33] propose a model to predict reaction time in response to natural images. This model is based on statistical properties of natural images and is claimed to accurately predict human reaction time by forming an entropy feature vector. Ratcliff et al. [34] used a drift diffusion model whose drift rate (the rate of accumulation of evidence towards a criterion) was determined by the quality of information to explain lexical decision times and accuracy (i.e. how rapidly does a person classify stimuli as words or non-words). Reaction time has also been studied in the context of perceptual decision making [3, 35, 36, 37]. Neural networks have been used to model object recognition [38], temporal dynamics in the brain [39, 40], the ventral stream, i.e., the object recognition neural pathway in human cortex [41], and temporal information [42]. Close to our approach, Spoerer et al. [18] has a similar goal of using a specific class of neural networks to model human reaction times, and are the first to use a neural network as a computational model of the speed-accuracy tradeoff. This work poses a binary classification problem (“animate” vs “inanimate” objects) to human observers and networks. However, a binary classification task may not represent general categorization accuracy because in a binary task an observer may learn to detect the difference between classes rather than actually classify images into classes. We discuss this approach in Section 4 and compare to it quantitatively in Section 5.

3 Collecting Behavioral Data

We measure accuracy and reaction time for human observers performing an object recognition task on images presented with and without perturbation. We assess the impact of adding color, blur, and noise, The results show a speed-accuracy tradeoff (Figure 1) for all three image manipulations. In

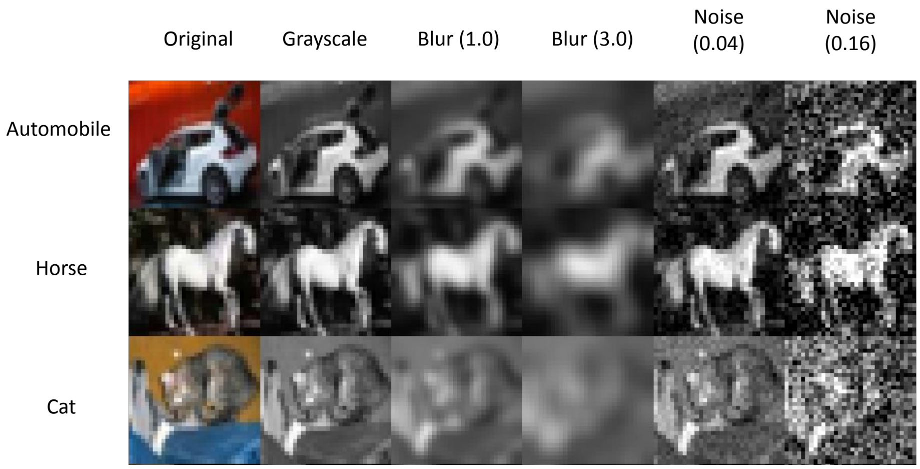


Figure 2: *Sample CIFAR-10 dataset [1] images are shown as originals (in color) on the left, along with our image perturbations (blur and noise addition), added to grayscale images.* The image perturbations provide control of the recognition task difficulty. Numbers in parentheses correspond to standard deviations for 0-mean Gaussian distributions. Units are pixels.

Sections 4 and 5, we evaluate the ability of neural networks to model the tradeoff between processing speed and accuracy. Our experimental protocol is similar to that of [32] and is outlined below.

Images. In all experiments, human observers recognized objects in CIFAR-10 images [1], a popular benchmark for neural network analysis, with the default train/test split. This image set contains 50,000 training images and 10,000 test images each of 32×32 pixels, and has 10 classes: airplane, automobile, bird, cat, deer, dog, frog, horse, ship and truck. Sample images and added perturbations can be seen in Figure 2. Unlike Spoerer et al. [18], we chose to analyze the CIFAR-10 images due to the smaller number of classes than in the ImageNet [4] dataset, making it easier for our human participants to memorize the relevant letter-key pairings to input responses. We used lab.js [43] and Just Another Tool for Online Studies (JATOS) [44] to present images and collect timed responses from human observers online. This software reliably gives accurate timing in benchmark evaluation of online testing packages [45], better than 5 ms trial-to-trial variation in stimulus duration and better than 10 ms trial-to-trial variation in reaction time, across many operating systems and browsers. Images were interpolated to 190×190 pixels for optimal viewing [46]. We estimate the size in cm of the 190x190 pixel image to be 4x4 cm, subtending 4x4 deg, and the viewing distance (distance between observer eye and screen) to be roughly 57 cm.

Observer statistics and data collection. We collected data from 142 observers (84 Male, 57 Female, 1 Non-binary) ranging in age from 24 to 62 years. Each session (set of trials) lasted about an hour. Each observer had a normal or corrected-to-normal vision. The stimuli were presented via JATOS survey via worker links to each observer. Participants were recruited through Amazon MTurk (similar to studies in [47, 48]), and paid $15 for their efforts (to a total of $2130 with all fees). A standard IRB approved (IRB-FY2016-404) consent form was signed before collecting the data by each observer, and demographic information was collected.

Table 1: *Summary statistics of collected data on human observers across all experiments.*

Exp.	#Partic.	Compl. time (min.)		#Trials
		Mean	SD	
Noise	57	50.80	17.47	1500
Blur	40	47.09	11.58	1500
Col. & Gr.	45	20.09	6.93	500

Survey design. The survey was designed to control the response time of human observers by asking them to respond in the allotted time distribution. The design was based on the previous work by McElree & Carrasco [32], where 4 observers participated in a total of 20 approximately 75 min sessions. At the beginning of each session subjects were instructed that each object category was linked to a particular letter-key: *(A)irplane, a(U)tomobile, (B)ird, (C)at, d(E)er, (D)og, (F)rog, (H)orse, (S)hip* and *(T)ruck*. They were then given a training run of 20 images where they learned the key-class labels and got feedback on the speed of their responses. A trial consisted of a stimulus image displayed for a fixed amount of time. Since 150 ms is the minimum visual processing time needed to process (recognize) a stimulus [49], the survey was designed on five fixed viewing conditions (blocks) at 200 ms, 400 ms, 600 ms, 800 ms, and 1000 ms with a tolerance of ±100 ms. Outside of these tolerance values, trials were discarded.

Note. The Carrasco & McElree speed-accuracy-tradeoff (SAT) paradigm [32] was a major advance in tracking the improvement of accuracy with time. However, allowing observers to respond when they feel like and then sorting into bins produces confounds that make the data hard to analyze because observers tend to take longer on harder trials. In our case, we trained observers to respond at a fixed time (different in each block), so measured accuracy is not confounded with trial-by-trial difficulty. Our use of their paradigm makes our results much easier to analyze. In many studies of the effect of timing in object recognition [48, 50, 51, 17], each trial's stimulus presentation and choice selection are separate steps. Various stimulus durations are reported: 100-2000 ms in [48], 100 ms in [50], 25-150 ms in [51], 200 ms in [17], after which the observers are allowed to take as much time as needed to make their selection. In our experiments (the SAT paradigm), each trial was one step. The image stayed on until the observer responded by pressing a key. Thus, our reported reaction times include all the time between stimulus onset and key press. Our observers had very little time to respond, compared to typical object recognition studies, and as a result, their accuracy appears lower than of those from other studies. In our case, the lowest timing threshold was specifically restricted so that the human accuracy is near chance.

During the experimental session, observers were asked to place their hands on the keyboard while being aware of the ten identifiers (A: Airplane, C: Cat and so on). They were instructed to answer at the beep as fast as possible to fall into the tolerance bounds, were given feedback after every trial and were continuously presented with a progress counter. Pressing the spacebar presented the next stimulus. Before starting the actual survey for data collection, a tutorial of 20 images was displayed to make observers understand the key mapping and get used to the timing protocol. To reduce the length of each experimental session, each observer responded to a randomly selected subset of 1,000 images. This image set was divided into approximately equal chunks across different amounts of perturbation (noise and blur). Figure 1 plots sample human accuracy as a function of reaction time. At 1,000 ms, most observers had accuracies about 40% to 50%, except for a few outliers.

Observer accuracy variance. To capture variability in observer responses, for noise and blur surveys, each time condition block consisted of 300 trials (1500 trials in total) while the color survey had 100 trials (500 trials in total). At the end of the time-limit for a trial, a beep sounded within 60 ms of which the observer had to enter their category decision via key-press after which feedback was given: if they were quick, slow or perfect while pressing the key. In Figure 3, we study how well the observers followed instructions to respond within the required time interval, by analyzing a plot of reaction time vs trial number. It can be seen that the fractional error in reaction time increases as the task becomes more difficult, that is, as the required reaction time is lowered. We also observed little difference between data collected online in our present study and very similar data from experiments with in-person testing [52].

4 Modelling Speed-Accuracy Tradeoff (SAT) with Networks

In order to test the ability of neural networks to capture the flexible, adaptive computation that humans exhibit, we analyze three representative models from existing literature. The first two, MSDNet [21] and SCAN [20], both state-of-the-art dynamic depth networks, were originally developed to improve test-time efficiency in computer vision applications. They are promising candidates for our purpose since they are capable of adaptive computation. We compare them against rCNN (which we refer to as ConvRNN) [18], a convolutional recurrent network which was recently developed specifically as a model for human speed-accuracy tradeoffs. It should be noted that, due to prior knowledge in humans and other confounding factors, it is difficult to replicate exactly the same training and testing conditions in humans and machines. To partially account for this, we perform trial runs for humans on sample data (see Section 3) and test both humans and networks on a variety of perturbation types and strengths. We compare networks with humans, first on *accuracy ranges* networks can achieve by only varying FLOPS used. Next, we measure networks' *correlation* with human behavior under various perturbation conditions, to determine if these models can capture the same performance trends that humans exhibit. Training details of each model are described in Supplementary Material.

- **Convolutional Recurrent Neural Network (ConvRNN)** ConvRNN [18] exhibits temporal behavior by relying on recurrent connectivity, characteristic of the primate visual system, implemented

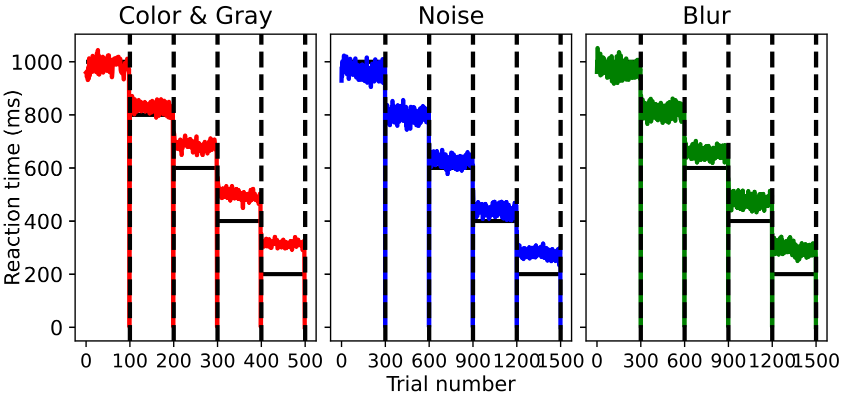


Figure 3: *The tight distribution of reaction times in each block of trials, across 3 experiments*. Plots show the reaction time (in ms) for each trial (averaged across observers) in our experiments. For each block, the black horizontal line denotes the beep timing at which the observer needs to respond. For analysis, we use the time of response, not the time of the beep, so the small SD of reaction time is good, and its delay relative to the beep is immaterial. Blocks are separated by vertical dotted lines.

by adding bottom-up and lateral connections to a feed-forward convolutional network. Lateral connections add cycles inside the feed-forward connectivity allowing for recurrent behavior. This model consists of 7 blocks of recurrent convolutional layers (RCL), followed by a Readout layer to output class predictions. During inference for a given image, the computation used by the model can be dynamically chosen by running the model for a variable number of recurrent cycles. This property allows the network to respond to an input image with a different amount of computation, which we use to represent reaction time.

- **Multi-Scale Dense Network (MSDNet)** MSDNet [21] implements dynamic inference using multiple early exit classifiers from a feedforward network. Since the exits are all at different depths in the network, classification at each one has a different computational requirement. All exits are placed after blocks of layers and use features from a common backbone network for classification. A consequence of this is that features deemed useful for each classifier during training interfere with the other classifiers. To resolve this problem, MSDNet proposes two architectural features: multi-scale feature maps, and dense connectivity (realized by using a DenseNet [53] backbone). These properties allow neurons at any layer to access features from any part of the network and at any resolution, thus diminishing the effect of the interference problem. In our experiments, we the number of scales, the bottleneck factor, and use a 15-layer backbone network with seven early exit classifiers placed at block intervals of 1-2-4, thus making up a total of seven blocks.
- **Scalable Neural Network (SCAN)** Similarly MSDNet, SCAN [20] implements dynamic inference using early exit classifiers from a common backbone network. Whereas MSDNet uses multi-scale feature maps and dense connectivity to solve the issue of interference between early and late classifiers, SCAN uses an encoder-decoder attention mechanism in each exit network. This allows each exit to "focus" only on features relevant for classification at a specific depth of the backbone. The attention network produces a binary mask which is added to the backbone (ResNet[54]) feature map, after which a Softmax layer predicts a class label. The network uses four early exits and a final ensemble output which uses all early exit features for prediction. Thus, for a given input, the network outputs five class predictions, each requiring a different amount of computation time/effort.

4.1 Contrast reduction

In our experiments on CIFAR-10 [1], networks were generally more accurate than human observers when presented with images (see Section 5) of the same noise levels. To match human accuracy, we added more noise to the images presented to networks, which resulted in noise levels that fell outside of the image pixel distribution. To avoid overflow without introducing clipping, we reduced the image contrast tenfold

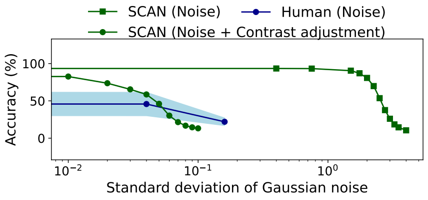


Figure 4: *Contrast reduction to avoid clipping.* So that the noisy image would not exceed the pixel range, we reduced contrast tenfold. This allows us to **a.** bring network accuracy to the same range as human accuracy, and **b.** increase task difficulty.

(see Figure 2 for examples). Contrast reduction removed the need for image clipping and brought the neural network accuracy closer to those produced by human observers. In Figure 4, we compared the accuracy of SCAN [20], the top performing network, on original and contrast-reduced images with and without noise, and found that the latter produced more human-like responses in networks than the former.

5 Results and discussion

We now study how well human response patterns are matched with results from our computational models. Specifically, we analyze the accuracy ranges exhibited by each model type and correlate model accuracy with human response slopes.

Comparing accuracy ranges. We analyze and compare human and neural network accuracy ranges. In Figure 5, we show the range of accuracies shown by each model and the human average. We find that the accuracies achieved by all networks greatly exceed that of human observers (by greater than 15%). On the other hand, the *accuracy range* (i.e., difference between maximum and minimum accuracies) is much higher in humans (51.37%) than in networks. Across the neural network models, MSDNet [21] offered the highest accuracy range (13.87%), followed by ConvRNN [18] (9.02%), and finally, SCAN [20] (4.34%). The large difference in accuracy range between humans and networks is primarily because networks achieve high classification accuracies even with low computational effort. Larger accuracy ranges can therefore be obtained by reframing the task to make it more challenging.

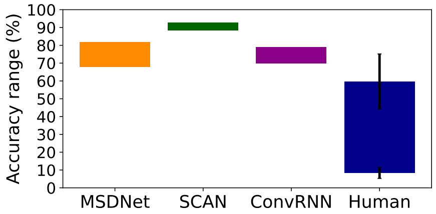


Figure 5: *Accuracy ranges of neural networks and a comparison with human observers.* The neural networks exhibit higher accuracies and smaller accuracy ranges than human observers.

Varying task difficulty using image perturbations. Here, we explore how well machines can adapt to task difficulty by comparing the accuracy range of both humans and networks on perturbed images. Noise in perception experiments is used for assessing unpredictable variation in some aspect of stimulus [55], and we attempt to model the same effect in our experiments. We modify the recognition task by adding noise and blur to make it more challenging, and then analyze the effect. Image perturbations are useful for bench-marking human accuracy [56, 57]. Additionally, CIFAR-10 is a relatively simple dataset on which the networks we considered performed well even at the most constrained settings (see Section 5). We therefore adjusted task difficulty by adding noise and blur to images, and retraining models with the perturbed training set. Figure 6 shows MSDNet's tradeoff curves under various amounts of test-time image noise. It can be seen that at zero noise, lowering computation below the lowest possible number of FLOPS would result in a catastrophic drop from 60% to chance. This is unlike humans whose accuracy drops more gracefully as allowed reaction time is lowered (Figure 1). Note that we conduct experiments primarily with grayscale images. We report the effect of color on human and network accuracy in Figure 7. We found that color improves the recognition accuracy for both humans and neural networks by only about 5% in both cases, and produced similar accuracy range patterns.

Finally, we correlate network accuracy to average human accuracy at varying levels of noise or blur, and report Pearson's r correlation coefficients in Figure 8a. To obtain an upper bound on correlations, we also correlate each human observer to the average human observer. Unlike previous work [18], which correlates only reaction time of humans and models, we report the correlation (Pearson's r) of model accuracies across different FLOPS levels with human accuracies across different time budgets. This metric captures both accuracy and reaction time and hence allows for a more robust evaluation of the speed-accuracy tradeoff exhibited by humans and models.

For blur, we find that the MSDNet [21] achieves the highest correlation to humans, followed by ConvRNN [18] and SCAN [20] while for noise, correlations of MSDNet and ConvRNN are both similar and higher than of SCAN. When comparing SCAN [20] models with different backbones, we find that decreasing the ResNet [54] backbone to ResNet-9 decreases the correlation. Similarly, choosing an over-parametrized ResNet-34 also adversely affects correlation (see Supplementary Material). It is important to point out the need for much higher noise to bring the network accuracy

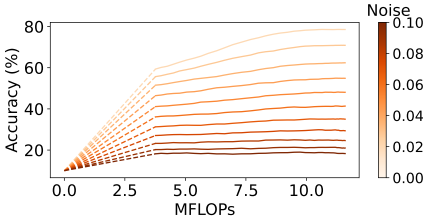


Figure 6: *MSDNet accuracy vs MFLOPS for various values of added image noise SD.* Each curve corresponds to a different Gaussian noise std. dev. $\in [0, 0.1]$, as shown by the color bar. Accuracy at 0 MFLOPS is taken to be at chance (10%), attained by any fixed response, and dotted lines extrapolate measured data points to this value. The dotted lines bridge the catastrophic failure of MSDNet, which cannot provide any useful answer with less than about 3.5 MFLOPS. The model is trained with random Gaussian noise of std. dev. $\in [0, 0.05]$.

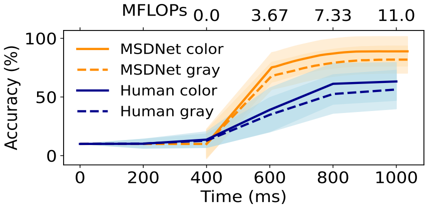


Figure 7: *Evaluation of the effect of color on network and human accuracy.* Color does not significantly affect the recognition accuracy of either humans or the MSDNet [21] model. Accuracy at 0 Time/MFLOPS was not measured and is assumed to be at chance. Linear transformation mapped MFLOPS (F) to Time (T) $[F = 11/600(T - 400)]$. Bounding areas represent standard deviations across observers.

down to the human level. This indicates that the neural networks are more tolerant to noise than human observers, once trained with noisy images.

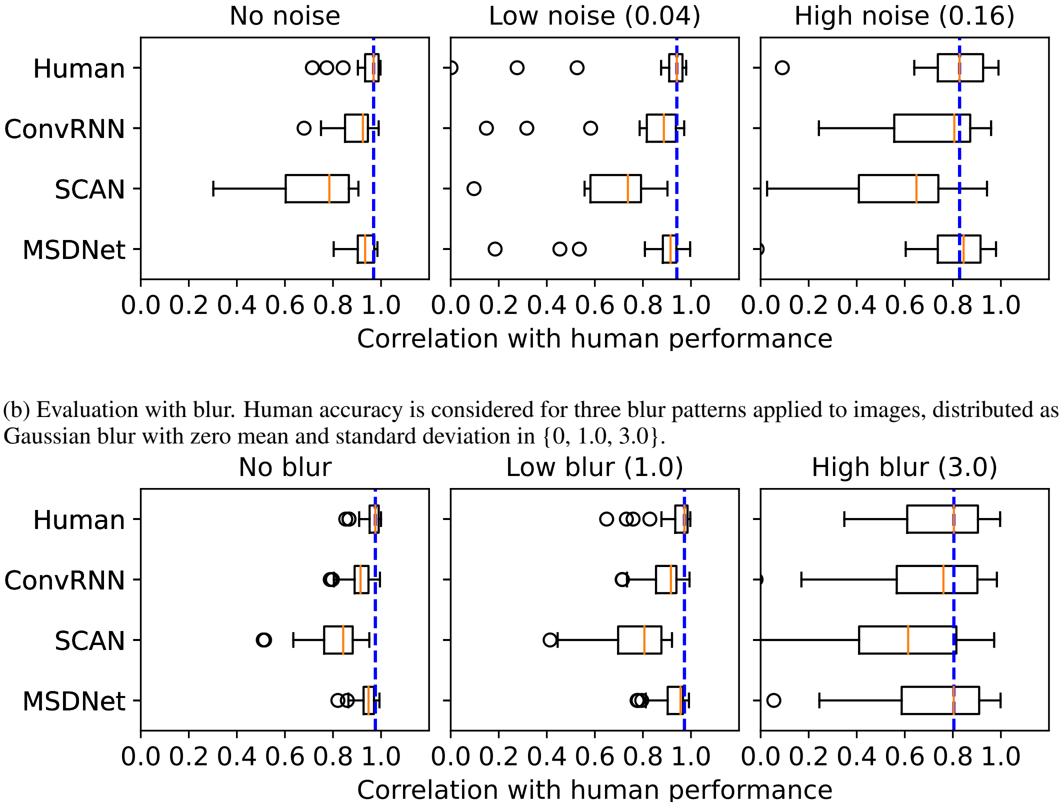
Table 2: *Pairwise comparison of models in terms of their correlation with human data indicates that most differences in correlation to humans observed in Figure 8 are statistically significant (p < 0.05).* Results were obtained using paired 2-tailed t-tests with Bonferroni correction. Model pairs with check marks indicate significant difference in correlation at a particular perturbation level.

Model Pair	Human Noise SD			Human Blur SD		
	0.0	0.04	0.16	0.0	1.0	3.0
	Significant difference?					
MSDNet - SCAN	✓	✓	✓	✓	✓	✓
MSDNet - ConvRNN	✓	✓	✓	✓	✓	✓
MSDNet - Human	✓	✓		✓	✓	
SCAN - ConvRNN	✓	✓	✓	✓	✓	✓
SCAN - Human	✓	✓	✓	✓	✓	✓
ConvRNN - Human	✓	✓	✓	✓	✓	

In Table 2, we report significance analysis of the above results, using paired 2-tailed t-tests, with Bonferroni correction to correct for multiple comparisons. A sample-size determination test showed that our sample size is large enough to draw all the above-mentioned conclusions [58]. Our analysis indicates that all of the comparison results discussed above are significant ($p \leq 0.05$, corrected). The comparisons for which the t-test indicated non-significant correlation are MSDNet - Human (high noise and blur) and ConvRNN - Human (high blur). Given that these noise and blur values present a nearly impossible case for human observers, this is a reasonable finding.

Extension to higher-resolution images. To evaluate the effect of increased resolution, we repeated our human and network experiments on STL-10 [2], a popular image recognition dataset. The STL-10 images each have 96x96 pixels, where the CIFAR-10 images have only 32x32. STL-10 has 10 classes of image, like CIFAR-10, but has only 800 images per class, where CIFAR-10 has 6,000. SAT results on STL-10 are presented in Figure 9*(a)*. We find that, with increased resolution, human accuracy improves, reaching a plateau of 70-75%, in comparison to 50-60% for CIFAR-10 (see Figure 1. This is consistent with earlier findings that people have trouble recognizing (low resolution) blocky images [59]. Corresponding correlations with the MSDNet network are shown in Figure 9*(b)*, and show that the network performance correlates well with human results, and that the difference in correlation is not statistically significant. However, in this experiment, the accuracy range of MSDNet is small compared to human observers, see Figure 9*(b)*. This finding shows that even though MSDNET

(a) Evaluation with noise. Human accuracy is considered for three noise patterns applied to images, distributed as Gaussian noise with zero mean and standard deviation in {0, 0.04, 0.16}.



(b) Evaluation with blur. Human accuracy is considered for three blur patterns applied to images, distributed as Gaussian blur with zero mean and standard deviation in {0, 1.0, 3.0}.

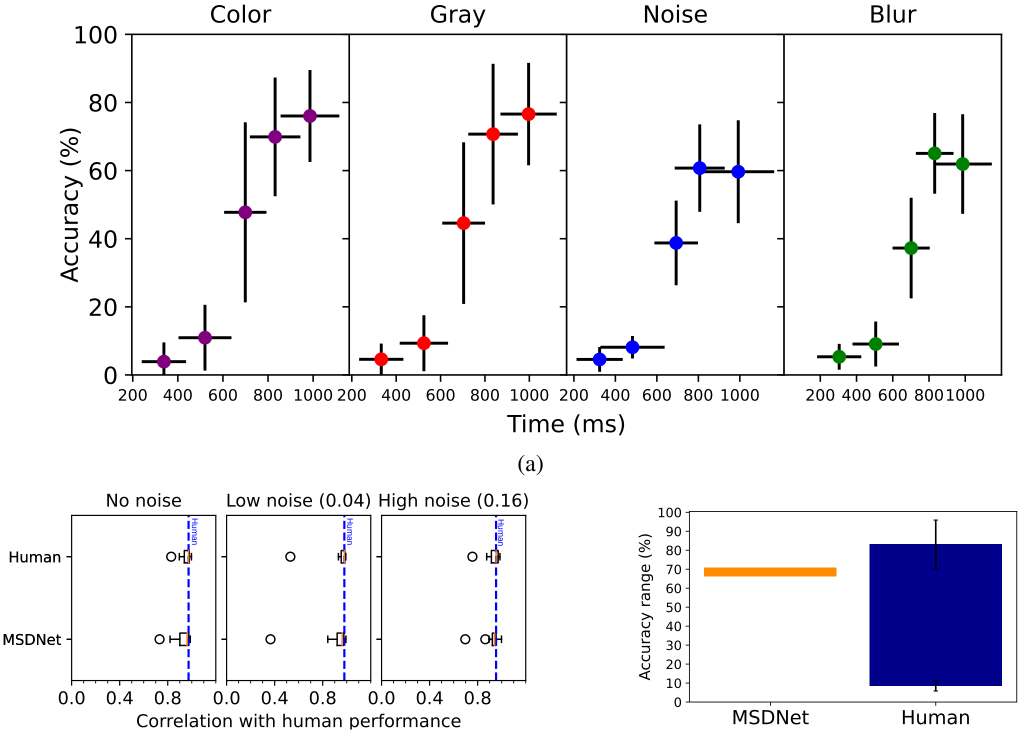
Figure 8: *Correlation of network accuracy with human accuracy across network FLOPS and human reaction time, respectively, evaluated at several levels of noise & blur.* For a fair comparison, the level of perturbation used during training is the same across all networks. During inference for each network, the noise or blur level that elicits the highest correlation with humans is found and shown above. MSDNet achieves the highest correlation with human observers in all testing scenarios. Orange bars represent median correlation value. Vertical blue line is an extension of the median correlation of humans with each other. Standard deviation for all correlations is shown.

achieved a high correlation, there remains a large gap in absolute performance between networks and humans.

6 Conclusion

Speed-accuracy tradeoff is an essential feature of human performance that is difficult to explain with current computational models of object recognition. We present a benchmark for timed object recognition by human observers, documenting their speed-accuracy tradeoff. We assess performance of several networks as models of SAT, and find that dynamic-depth neural networks are promising. To compare various networks with humans, we propose two metrics: (1) accuracy range and (2) the correlation of SAT between network and humans. The two metrics capture the magnitude of the model's SAT and its similarity to the human SAT.

One of the considered models, MSDNet [21], gives a better account than previous attempts [18], without the need for recurrence. In the presence of noise or blur, MSDNet accuracy deteriorates much as human accuracy does. When trained with noise (or blur), it shows a 0.93 (or 0.94) correlation with human accuracy. Finally, we test the effect of network-backbone architecture and determine that correlation to human accuracy typically does not increase with additional parameters.



(a)

(b)

(c)

Figure 9: *Results on STL-10 [2] dataset.* (a) Mean and standard deviation of accuracy and reaction time across participants dataset. (b) Correlation between MSDNet and human performance. (c) Accuracy range of MSDNet and humans.

While dynamic networks succeed in showing some speed-accuracy tradeoff, they achieve a much smaller range than humans do. The average human accuracy range is 51% while the best network, MSDNet trained with noise, achieves only a 19% range. With high perturbation, humans stumble and machines fall. This motivates future work that aims to build neural networks that can better match the flexibility and adaptability of human object recognition. Work in this direction is important in understanding human decision making and deploying machine-learning in time-sensitive applications.

Applications of the above-described technology have potential benefits (addressing public health concerns — e.g. slow reading — and biases in computational models) and risks (facilitating the creation of bots that pass for humans for malicious purposes). Such concerns are shared by much research in computational modeling, and are outside the scope of this work.

References

[1] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

[2] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 215–223. JMLR Workshop and Conference Proceedings, 2011.

[3] John Palmer, Alexander C Huk, and Michael N Shadlen. The effect of stimulus strength on the speed and accuracy of a perceptual decision. *Journal of vision*, 5(5):1–1, 2005.

[4] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 2015.

[5] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[6] Ian Lenz, Honglak Lee, and Ashutosh Saxena. Deep learning for detecting robotic grasps. *The International Journal of Robotics Research*, 34(4-5):705–724, 2015.

[7] Oludare Isaac Abiodun, Aman Jantan, Abiodun Esther Omolara, Kemi Victoria Dada, Nachaat AbdElatif Mohamed, and Humaira Arshad. State-of-the-art in artificial neural network applications: A survey. *Heliyon*, 4(11):e00938, 2018.

[8] Yizeng Han, Gao Huang, Shiji Song, Le Yang, Honghui Wang, and Yulin Wang. Dynamic neural networks: A survey. *arXiv preprint arXiv:2102.04906*, 2021.

[9] Mengwei Xu, Feng Qian, Mengze Zhu, Feifan Huang, Saumay Pushp, and Xuanzhe Liu. Deepwear: Adaptive local offloading for on-wearable deep learning. *IEEE Transactions on Mobile Computing*, 19(2):314–330, 2019.

[10] Ming Yang, Shige Wang, Joshua Bakita, Thanh Vu, F Donelson Smith, James H Anderson, and Jan-Michael Frahm. Re-thinking cnn frameworks for time-sensitive autonomous-driving applications: Addressing an industrial challenge. In *2019 IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS)*, pages 305–317. IEEE, 2019.

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

[12] Kunihiko Fukushima and Sei Miyake. Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In *Competition and cooperation in neural nets*, pages 267–285. Springer, 1982.

[13] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144, 2018.

[14] Anthony M Zador. A critique of pure learning and what artificial neural networks can learn from animal brains. *Nature communications*, 10(1):1–7, 2019.

[15] Joel Dapello, Tiago Marques, Martin Schrimpf, Franziska Geiger, David D Cox, and James J DiCarlo. Simulating a primary visual cortex at the front of cnns improves robustness to image perturbations. In *NeurIPS*, 2020.

[16] Michael L Iuzzolino, Michael C Mozer, and Samy Bengio. Improving anytime prediction with parallel cascaded networks and a temporal-difference loss. *arXiv preprint arXiv:2102.09808*, 2021.

[17] Robert Geirhos, Carlos R Medina Temme, Jonas Rauber, Heiko H Schütt, Matthias Bethge, and Felix A Wichmann. Generalisation in humans and deep neural networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 7549–7561, 2018.

[18] Courtney J Spoerer, Tim C Kietzmann, Johannes Mehrer, Ian Charest, and Nikolaus Kriegeskorte. Recurrent neural networks can explain flexible trading of speed and accuracy in biological vision. *PLoS computational biology*, 16(10):e1008215, 2020.

[19] Weizhe Hua, Yuan Zhou, Christopher De Sa, Zhiru Zhang, and G Edward Suh. Channel gating neural networks. *arXiv preprint arXiv:1805.12549*, 2018.

[20] Linfeng Zhang, Zhanhong Tan, Jiebo Song, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. SCAN: A scalable neural networks framework towards compact and efficient models. *NeurIPS*, 2019.

[21] Gao Huang, Danlu Chen, Tianhong Li, Felix Wu, Laurens van der Maaten, and Kilian Weinberger. Multi-scale dense networks for resource efficient image classification. In *ICLR*, 2018.

[22] Demis Hassabis, Dharshan Kumaran, Christopher Summerfield, and Matthew Botvinick. Neuroscience-inspired artificial intelligence. *Neuron*, 95(2):245–258, 2017.

[23] Anthony J Bell and Terrence J Sejnowski. The "independent components" of natural scenes are edge filters. *Vision research*, 37(23):3327–3338, 1997.

[24] Grace W Lindsay. Attention in psychology, neuroscience, and machine learning. *Frontiers in computational neuroscience*, 14:29, 2020.

[25] Xitong Gao, Yiren Zhao, Łukasz Dudziak, Robert Mullins, and Cheng-zhong Xu. Dynamic channel pruning: Feature boosting and suppression. *arXiv preprint arXiv:1810.05331*, 2018.

[26] Zhourong Chen, Yang Li, and Si Si Bengio, Sami. You look twice: Gaternet for dynamic filter selection in cnns. *CVPR*, 2019.

[27] Changlin Li, Guangrun Wang, Bing Wang, Xiaodan Liang, Zhihui Li, and Xiaojun Chang. Dynamic slimmable network. *arXiv preprint arXiv:2103.13258*, 2021.

[28] Tolga Bolukbasi, Joseph Wang, Ofer Dekel, and Venkatesh Saligrama. Adaptive neural networks for fast test-time prediction. *arXiv preprint arXiv:1702.07811*, 2017.

[29] Xin Wang, Fisher Yu, Zi-Yi Dou, Trevor Darrell, and Joseph E Gonzalez. Skipnet: Learning dynamic routing in convolutional networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 409–424, 2018.

[30] Andreas Veit and Serge Belongie. Convolutional networks with adaptive inference graphs. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–18, 2018.

[31] Christopher Summerfield and Floris P De Lange. Expectation in perceptual decision making: neural and computational mechanisms. *Nature Reviews Neuroscience*, 15(11):745–756, 2014.

[32] B. McElree and M. Carrasco. The temporal dynamics of visual search: evidence for parallel processing in feature and conjunction searches. *J Exp Psychol Hum Percept Perform*, 1999.

[33] A. Mirzaei, S. M. Khaligh-Razavi, M. Ghodrati, S. Zabbah, and R. Ebrahimpour. Predicting the human reaction time based on natural image statistics in a rapid categorization task. *Vision Res.*, 2013.

[34] R. Ratcliff, P. Gomez, and G. McKoon. A diffusion model account of the lexical decision task. *Psychol Rev*, 2004.

[35] Kong-Fatt Wong and Xiao-Jing Wang. A recurrent network mechanism of time integration in perceptual decisions. *Journal of Neuroscience*, 2006.

[36] Eric-Jan Wagenmakers, Han LJ Van Der Maas, and Raoul PPP Grasman. An EZ-diffusion model for response time and accuracy. *Psychonomic bulletin & review*, 2007.

[37] Ulrike Basten, Guido Biele, Hauke R Heekeren, and Christian J Fiebach. How the brain integrates costs and benefits during decision making. *Proceedings of the National Academy of Sciences*, 2010.

[38] Courtney J Spoerer, Patrick McClure, and Nikolaus Kriegeskorte. Recurrent convolutional neural networks: a better model of biological object recognition. *Frontiers in psychology*, 8:1551, 2017.

[39] Tim C Kietzmann, Courtney J Spoerer, Lynn KA Sörensen, Radoslaw M Cichy, Olaf Hauk, and Nikolaus Kriegeskorte. Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences*, 116(43):21854–21863, 2019.

[40] Umut Güçlü and Marcel AJ van Gerven. Modeling the dynamics of human brain activity with recurrent neural networks. *Frontiers in computational neuroscience*, 11:7, 2017.

[41] Qianli Liao and Tomaso Poggio. Bridging the gaps between residual learning, recurrent neural networks and visual cortex. *arXiv preprint arXiv:1604.03640*, 2016.

[42] Zedong Bi and Changsong Zhou. Understanding the computation of time using neural network models. *Proceedings of the National Academy of Sciences*, 117(19):10530–10540, 2020.

[43] Felix Henninger, Yury Shevchenko, Ulf Mertens, Pascal J. Kieslich, and Benjamin E. Hilbig. lab.js: A free, open, online experiment builder, July 2020.

[44] Kristian Lange, Simone Kühn, and Elisa Filevich. "just another tool for online studies” (JATOS): An easy solution for setup and management of web servers supporting online studies. *PLOS ONE*, 10(6), jun 2015.

[45] David Bridges, Alain Pitiot, Michael R MacAskill, and Jonathan W Peirce. The timing mega-study: Comparing a range of experiment generators, both lab-based and online. *PeerJ*, 8:e9414, 2020.

[46] D. G. Pelli. VISUAL SCIENCE:close encounters–an artist shows that size affects shape. *Science*, 285(5429):844–846, aug 1999.

[47] Jeremy Freeman, Corey M Ziemba, David J Heeger, Eero P Simoncelli, and J Anthony Movshon. A functional and perceptual signature of the second visual area in primates. *Nature neuroscience*, 16(7):974–981, 2013.

[48] Najib J Majaj, Ha Hong, Ethan A Solomon, and James J DiCarlo. Simple learned weighted sums of inferior temporal neuronal firing rates accurately predict human core object recognition performance. *Journal of Neuroscience*, 35(39):13402–13418, 2015.

[49] Simon Thorpe, Denis Fize, and Catherine Marlot. Speed of processing in the human visual system. *Nature*, 381(6582):520–522, jun 1996.

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

[51] Hanlin Tang, Martin Schrimpf, William Lotter, Charlotte Moerman, Ana Paredes, Josue Ortega Caro, Walter Hardesty, David Cox, and Gabriel Kreiman. Recurrent computations for visual pattern completion. *Proceedings of the National Academy of Sciences*, 115(35):8835–8840, 2018.

[52] Omkar Kumbhar, Elena Sizikova, Najib Majaj, and Denis G Pelli. Anytime prediction as a model of human reaction time. *arXiv preprint arXiv:2011.12859*, 2020.

[53] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *CVPR*, 2017.

[54] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[55] Remy Allard, Jocelyn Faubert, and Denis G. Pelli. Editorial: Using noise to characterize vision. *Frontiers in Psychology*, 2015.

[56] Ineke MCJ van Overveld. Contrast, noise, and blur affect performance and appreciation of digital radiographs. *Journal of digital imaging*, 8(4):168, 1995.

[57] Denis G Pelli and Bart Farell. Why use noise? *JOSA A*, 16(3):647–653, 1999.

[58] S Hulley, S Cummings, W Browner, D Grady, and T Newman. Designing clinical research (vol. 4th). *Philadelphia: LWW*, 2013.

[59] Denis G Pelli. Close encounters–an artist shows that size affects shape. *Science*, 285(5429):844–846, 1999.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See Section 1.
   (b) Did you describe the limitations of your work? [Yes] See Section 6.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Section 6.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes] See Section 6 and Supplementary Material.
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)?
   [Yes] See `https://osf.io/zkvep/` for dataset. and `https://github.com/ajaysub110/anytime-prediction` for code.
   (b) Did you specify all the training details (e.g., data splits, hyper-parameters, how they were chosen)? [Yes] See Section 4.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Section 5.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Supplementary Material.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See Section 3.
   (b) Did you mention the license of the assets? [Yes] See Section 3.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] See Section 3.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] See Section 3.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] See Section 3.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [Yes] See website https://github.com/ajaysub110/anytime-prediction.
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [Yes] See website https://github.com/ajaysub110/anytime-prediction.
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [Yes] See Section 3.