PsyPhy: A Psychophysics Driven Evaluation Framework for Visual Recognition

Brandon RichardWebster, *Student Member, IEEE,*
Samuel E. Anthony, *Student Member, IEEE,*
and Walter J. Scheirer, *Senior Member, IEEE*

***Abstract*—By providing substantial amounts of data and standardized evaluation protocols, datasets in computer vision have helped fuel advances across all areas of visual recognition. But even in light of breakthrough results on recent benchmarks, it is still fair to ask if our recognition algorithms are doing as well as we think they are. The vision sciences at large make use of a very different evaluation regime known as Visual Psychophysics to study visual perception. Psychophysics is the quantitative examination of the relationships between controlled stimuli and the behavioral responses they elicit in experimental test subjects. Instead of using summary statistics to gauge performance, psychophysics directs us to construct item-response curves made up of individual stimulus responses to find perceptual thresholds, thus allowing one to identify the exact point at which a subject can no longer reliably recognize the stimulus class. In this article, we introduce a comprehensive evaluation framework for visual recognition models that is underpinned by this methodology. Over millions of procedurally rendered 3D scenes and 2D images, we compare the performance of well-known convolutional neural networks. Our results bring into question recent claims of human-like performance, and provide a path forward for correcting newly surfaced algorithmic deficiencies.**

***Index Terms*—Object Recognition, Visual Psychophysics, Neuroscience, Psychology, Evaluation, Deep Learning.**

I. Introduction

> We often attribute "understanding" and other cognitive predicates by metaphor and analogy to cars, adding machines, and other artifacts, but nothing is proved by such attributions.
>
> *John Searle*

Imagine the following scenario: a marvelous black box algorithm has appeared that purportedly solves visual object recognition with human-like ability. As a good scientist, how might you go about falsifying this claim? By all accounts, the algorithm achieves superior performance on established benchmarks in computer vision, but its internal workings are opaque to the external observer. Such a situation is not far fetched — it should be familiar to any of us studying machine learning for visual recognition. But what many of us in computer vision might not realize is that this setup happens to be the classic Chinese Room [3] problem proposed by the philosopher John Searle.

In Searle's thought experiment, a person who does not speak Chinese is alone in a locked room and following instructions from a computer program to generate Chinese characters to respond to Chinese messages that are slipped under the door. To the message passer outside of the room, the person inside understands Chinese. However, this is not the case. The person inside the room is simply following instructions to complete the task — there is no real replication of the competency of knowing the Chinese language. Linking this back to computer vision, the summary statistics of performance from our algorithms look good on benchmark tests — enough so that we believe them to be close to human performance in some cases. But are these algorithms really solving the general problem of visual object recognition, or are they simply leveraging "instructions" provided in the form of labeled training data to solve the dataset?

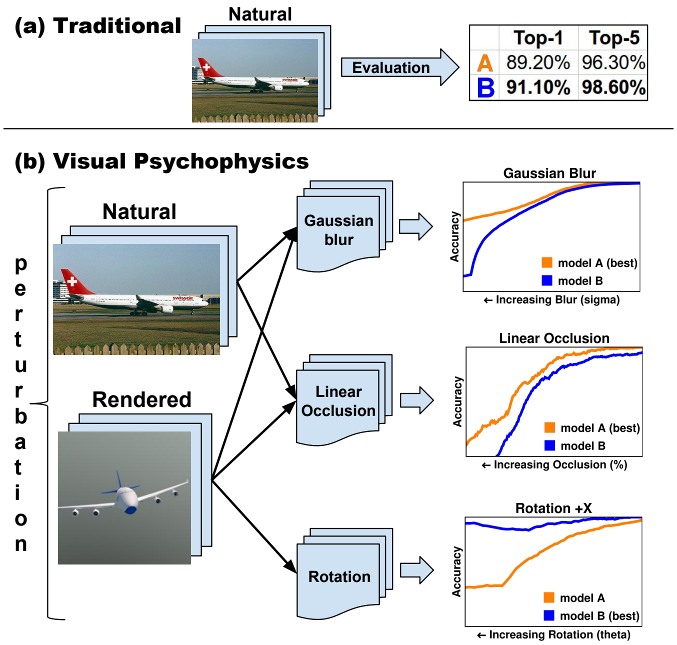

Fig. 1. In this article, the concept of applying psychophysics [1], [2] on a recognition model is introduced. In this figure, A and B are two models being compared. (Top) In traditional dataset-based evaluation, summary statistics are generated over large sets of data, with little consideration given to specific conditions that lead to incorrect recognition instances. (Bottom) Psychophysics, a set of experimental concepts and procedures from psychology and neuroscience, helps us plot the exact relationships between perturbed test images and resulting model behavior to determine the precise conditions under which models fail. Instead of comparing summary statistics, we compare item-response curves representing performance (y-axis) versus the dimension of the image being manipulated (x-axis).

Datasets in computer vision are intended to be controlled testbeds for algorithms, where the task and difficulty can be modulated to facilitate measurable progress in research. A dataset could be made up of images specifically acquired for experimentation, or publicly available images crawled from the web. Under this regime, strong advancements have been demonstrated for a number of problems, most notably for object recognition [4]. Deep learning is now a mainstay in computer vision thanks in part to the 2012 ImageNet Challenge [5], where AlexNet [6] reduced top-5 object classification error to 16.4% from the previously best reported result of 25.8%. When algorithms are evaluated on a common footing, it is possible to track meaningful improvements in artificial intelligence like this one. However, increases in error when different datasets are used for training and testing [7],

B. RichardWebster and W. Scheirer are with the Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN, 46556. Corresponding Author's E-mail: brichar1@nd.edu.

S. Anthony is with the Department of Psychology, Harvard University, and Perceptive Automata, Inc.

[8] make us wonder if this is the *only* way to do it.

When it comes to natural intelligence, neuroscientists and psychologists do not evaluate animals or people in the same way that computer vision scientists evaluate algorithms — and for a very good reason. With a collection of images crawled from the web, there is no straightforward way to determine the exact condition(s) that caused a subject to fail at recognizing a stimulus presented during an experiment. A natural image is the product of the physics at the instant the sensor acquired the scene; its latent parameters are largely unknown. Instead, for behavioral experiments meant to discover perceptual thresholds (*i.e.*, the average point at which subjects start to fail), the vision sciences outside of computer vision use the concepts and procedures from the discipline of *visual psychophysics*.

Psychophysics is the quantitative study of the relationships between controlled stimuli and the behavioral responses they elicit in a subject [1], [2]. It is a way to probe perceptual processes through the presentation of incremental and, in many cases, extremely fine-grained perturbations of visual stimuli. The properties of each stimulus are varied along one or more physical dimensions, thus controlling the difficulty of the task. The result (Fig. 1) is an *item-response curve* [9], where performance (*e.g.*, accuracy) on the y-axis is plotted against the dimension being manipulated (*e.g.*, Gaussian blur) on the x-axis. Each point on the curve reflects an individual stimulus, letting us map performance back to causal conditions in a precise manner. Psychophysics is an indispensable tool to vision science, and has been deployed to discover the minimum threshold for stimulation of a retinal photoreceptor (a single photon) [10], confirm Helmholtz's assertions on color absorption in the retina [11], and establish criteria to diagnose prosopagnosia [12] (the inability to recognize a face). As in these discoveries from biological vision, we submit that psychophysics holds much promise for discovering new aspects of the inner workings of machine learning models.

In this article, we introduce a comprehensive evaluation framework for visual recognition that is underpinned by the principles of psychophysics. In this regime, a stimulus can be an object drawn from purely rendered data or natural scene data, and a varying physical parameter can control the amount of transformation in the subsequent set of manipulated images derived from the original stimulus. A key difference from traditional benchmarks in computer vision is that instead of looking at summary statistics (*e.g.*, average accuracy, AUC, precision, recall) to compare algorithm performance, we compare the resulting item-response curves. For complete control of the underlying parameter space, we find that procedural graphics [13]–[16] are a useful way to generate stimuli that can be manipulated in any way we desire. Because we have the procedure that rendered each scene, we can find out where a model is failing at the parametric level. As we will see, by using this framework to explore artificial vision systems like psychologists, many interesting new findings can be surfaced about the strengths and limitations of computer vision models.

To summarize, our main contributions are as follows:

- A general evaluation framework is developed for performing visual psychophysics on computer vision models. The framework has a strong grounding in well-established work in psychology and neuroscience for behavioral experimentation.
- An investigation of procedural graphics for large-scale psychophysics experiments applied to models.
- A parallelized implementation of the psychophysics framework that is deployable as a Python package.
- A case study consisting of a battery of experiments incorporating millions of procedurally rendered images and 2D images that were perturbed, performed over a set of well-known Convolutional Neural Network (CNN) models [6], [17]–[19].

II. Related Work

Methods of Evaluation from the Vision Sciences. With respect to work in computer vision directly using psychophysics, most is related to establishing human baselines for comparison to algorithmic approaches. Riesenhuber and Poggio [20] described a series of psychophysical comparisons between humans and the HMAX [21] model of visual cortex using a limited set of stimuli rendered by computer graphics. Similarly, Eberhardt et al. [22] designed an experiment to measure human accuracy and reaction time during visual categorization tasks with natural images, which were then compared to different layers of CNN models [6], [19]. Geirhos et al. undertook a similar study for image degradations [23]. With respect to low-level features, Gerhard et al. [24] introduced a new psychophysical paradigm for comparing human and model sensitivity to local image regularities.

Psychophysics can also be used for more than just performance evaluation. Scheirer et al. [25] introduced the notion of "perceptual annotation" for machine learning, whereby psychophysical measurements are used as weights in a loss function to give a training regime some *a priori* notion of sample difficulty. Using accuracy and reaction time measured via the online psychophysics testing platform TestMyBrain.org [26], perceptual annotation was shown to enhance face detection performance. Along these lines, Vondrick et al. [27] devised a method inspired by psychophysics to estimate useful biases for recognition in computer vision feature spaces.

Outside of work specifically invoking psychophysics, one can find other related methods from psychology and neuroscience for behavioral-style model testing. 2D natural images are the most common type of data in computer vision, and form a good basis for algorithmic evaluation in this mode. O'Toole et al. [28], [29] and Philips et al. [30] have designed controlled datasets of natural images to compare human face recognition performance against algorithms. With the focus on algorithmic consistency with human behavior, there is no explicit model vs. model comparison in these methods.

More control in experimentation can be achieved through the use of rendered 3D scenes. Cadiue et al. [31], Yamins et al. [32] and Hong et al. [33] all make use of rendered images with parametrized variation to compare the representations of models with those found in the primate brain. Pramod and Arun [34] describe a set of perceived dissimilarity measurements from humans that is used to study the systematic differences between human perception and a large number of handcrafted and learned feature representations. Because of a need for very fine-grained control of object parts and other latent parameters of scenes, procedural graphics were introduced by Tenenbaum et al. [13] for the study of one-shot learning using probabilistic generative models. The use of procedural graphics for generative models was further developed by Yildirim et al. [14], Kulkarni et al. [15], and Wu et al. [16]. These studies do not vary the conditions of the stimuli using the procedures of psychophysics, nor do they use large-scale renderings on the order of millions of scenes.

Other Manipulations of Stimuli in Visual Recognition Evaluations. Work coming directly out of computer vision also addresses stimulus generation for the purpose of isolating model weaknesses. Hoiem et al. [35] suggest systematically varying occlusion, size, aspect ratio, visibility of parts, viewpoint, localization error, and background to identify errors in object detectors. Wilber et al. [36] systematically apply noise, blur, occlusion, compression, textures and warping effects over 2D scenes to assess face detection performance. Finally, a whole host of approaches can be found to manipulate the inputs to CNNs in order to highlight unexpected classification errors. These include the noise patterns introduced by Szegedy et al. [37] that are imperceptible to humans, and the fooling images produced via evolutionary algorithms that were explored by Nguyen et al. [38] and Bendale and Boult [39]. The level of control in the evaluation procedures varies between these approaches, but a common starting point based on model preference for each class is missing (*i.e.*, which object configuration produces the highest score?). We suggest in this article that the use of computer graphics helps us address this.

III. Method: The PsyPhy Framework

Our procedure for performing psychophysics on a model largely follows established procedures found in psychology, with a few key adaptations to accommodate artificial perception. For the purposes of this article, our focus is on two performance-based forced-choice tasks that yield an interpretable item-response curve. For descriptions of other procedures in psychophysics, see [1], [2]. First, let us consider the *two-alternative forced choice (2AFC) match-to-sample* task that is common in psychological testing.

In the 2AFC match-to-sample procedure, an observer is shown a "sample" stimulus, followed by two "alternate" stimuli where one is a positive (*i.e.*, matching) stimulus and the other is a negative (*i.e.*, non-matching) stimulus. The observer is then asked to choose from the alternate stimuli the stimulus that best matches the sample — the match criterion may or may not be provided to the observer. The observer repeats the task at different perturbed stimulus levels in either an adaptive pattern, which is like gradient descent for humans, or via the method of constants, which is a predetermined set of perturbed stimulus levels. Regardless of method, each task has two presented alternate stimuli ($N = 2$) and thus two-alternative forced-choices ($M = 2$). Analysis of the experiment would utilize the mean or median accuracy humans achieved at each stimulus level and mean or median human response time, if recorded. Models can be tested in precisely the same way when the input images are arranged as described above and accuracy is used as the performance measure.

Algorithm 1 $D_f^m(i,c)$, the top-1 binary decision of the Softmax layer of a CNN. Used for both preferred view calculation and MAFC.

```
Input: $f$, a single pre-trained network model
Input: $i$, an input image
Input: $c$, the expected class
$V = f(i)$                                  ▷ the Softmax vector
$c^* = \mathrm{argmax}_{j\in[0,|V|)} V_j$   ▷ find class label
$\varsigma = V_{c^*}$
if $c \neq c^*$ then                        ▷ incorrect class, negate response
    $\varsigma = -1 * \varsigma$
end if
return $\varsigma$, the decision score
```

Second, we can consider a mapping of a more difficult classification procedure in machine learning to a more general version of the 2AFC match-to-sample procedure. We call this mapped classification *MAFC match-to-sample*. In MAFC, the probe image in classification is equivalent to the sample stimulus. In classification, we rarely have only two classes for a model to choose from. Thus the value of M becomes the number of labeled training classes (*e.g.*, ImageNet 2012 has 1K learned classes, making $M = 1\text{K}$). Likewise, N — the number of presented alternate stimuli — changes to the number of images used in training, as this is the set of images the model is implicitly matching to (*e.g.*, for ImageNet 2012, $N = \sim 1.2\text{M}$ training images).

When testing a model with any psychophysics procedure, we need a special process for the selection of the stimuli's default state, that is, where there is no perturbation. Blanz et al. [40] show that humans have a natural inclination towards recognizing a certain view of an object, called a canonical view. We assume in human trials that an object configuration close to a canonical view will be chosen, maximizing the probability that all observers will have no problems performing at least some part of a match-to-sample task. However, this is not as simple for any task involving a model because we do not necessarily know if it follows a similar canonical view. But we can say that a model's *preferred view* is a view that produces the strongest positive response, as determined by a decision score. Note that there can be more than one preferred view (hence our use of the term *preferred*), because ties are often observed for the strongest response, especially in quantized decision score spaces. Choosing a preferred view is crucial to guaranteeing that when the stimulus is perturbed, the model's response will already be at its maximum for that class. Any perturbation will cause a decline (or possibly no change) in the strength of the model's response, not an increase.

PsyPhy Framework for 2AFC and MAFC. Inspired by software frameworks for subject testing like PsychoPy [41], we have implemented the 2AFC and MAFC procedures described above using a Python framework for model testing called PsyPhy. Here we describe the details of each component of the framework. The basic steps of (1) stimuli selection, (2) preferred view selection, (3) perturbation, and (4) item-response

Algorithm 2 $\phi_T^2(D_f^2, V, s)$: an item-response point generation function supporting 2AFC tasks for any image transformation function $T(s, v)$

```
Input: $D_f^2$, decision function for 2AFC
Input: $V$, a vector of preferred views for a set of classes
Input: $s$, the stimulus level
$h(v)$ :=random $v' | v' \in [V \setminus \{v\}]$    ▷ pick negative
$\beta = \sum_{v \in V} \max(0, \lceil D_f^2(T(s, v), v, h(v)) \rceil)$
$a = \frac{\beta}{|V|}$
return $\{s, a\}$, an $x, y$ coordinate pair (stimulus level, accuracy over trials) for one item-response point
```

Algorithm 3 $\phi_T^m(D_f^m, V, s)$: an item-response point generation function supporting MAFC tasks for any image transformation function $T(s, v)$

```
Input: $D_f^m$, decision function for MAFC
Input: $V$, a vector of preferred views for a set of classes
Input: $s$, the stimulus level
$\beta = \sum_{v \in V} \max(0, \lceil D_f^m(T(s, v), c(v)) \rceil)$
$a = \frac{\beta}{|V|}$
return $\{s, a\}$, an $x, y$ coordinate pair (stimulus level, accuracy over trials) for one item-response point
```

curve generation apply to any psychophysics procedure, and the specific 2AFC and MAFC procedures may be viewed as pluggable modules within the framework. PsyPhy is very flexible with respect to tasks it can support.

The first step is to select the initial set of stimuli for each class. For 2D natural images, this is any set of chosen images I_{2D} for a class c. For a rendered scene, a set of image specifications I_{3D} is provided to a rendering function $R(c, v)$ (implemented in this work using Mitsuba [42]) to render a single object centered in an image. The view $v \in I_{3D}$ is the parameter set $\{x, y, z, \psi\}$, where the coordinates x, y, and z are real numbers in the range $(-180.0, 180.0]$ and ψ, representing scale, is a real number in the range $(0.0, 25.0]$.

The second step is to find an individual model's preferred view for each class. For natural 2D images, the preferred view function in Eq. 1 is used. The second preferred view function, Eq. 2, uses R to create rendered images for classification. In Eq. 2, the search space is almost infinite, thus it does not find the absolute global maximum, but rather an approximation.

$$\mathcal{P}_{2D}(I_{2D}, c) := \operatorname*{argmax}_{i \in I_{2D}} D_f^m(i, c) \tag{1}$$

$$\mathcal{P}_{3D}(I_{3D}, c) := \operatorname*{argmax}_{v \in I_{3D}} D_f^m(R(c, v), c) \tag{2}$$

A decision function for classification $D_f^m(i, c)$ (Alg. 1) normalizes the score output of a model f to a value in the range $[-1.0, 1.0]$, which gives both a decision and a confidence associated with that decision. A value in the range $[-1.0, 0]$ is an incorrect decision and $(0, 1.0]$ is a correct decision. The parameter i is the input stimulus and c is the expected class.

A natural 2D preferred view (Eq. 1) is a single selected image $i \in I_{2D}$, where D_f^m has the strongest positive response. A 3D preferred view (Eq. 2) is a single selected set $v = \{x_v, y_v, z_v, \psi_v\} \in I_{3D}$, where D_f^m has the strongest positive response. The major difference between Eq. 1 and Eq. 2 is the use of R in Eq. 2 to render the image prior to measuring the response from D_f^m. Invoking Eq. 1 or Eq. 2 for each class will create a vector of preferred views V.

After preferred views have been selected for all classes, whether natural or rendered, the next step is to apply perturbations to them. In this procedure, a set of preferred views is perturbed at a specific stimulus level (*i.e.*, the amount of perturbation) using a function $T(s, v)$, where T could be any image transformation function (*e.g.*, Gaussian blur, rotation). The parameter v is one preferred view — either in 2D image format or $\{x, y, z, \psi\}$ for rendered stimuli — and s is the stimulus level. The function $\phi_T(D, V, s)$ perturbs the set of preferred views given in V and then makes a decision on each image using a decision function D. The specific implementation $\phi_T^2(D_f^2, V, s)$ for 2AFC is described in Alg. 2, and $\phi_T^m(D_f^m, V, s)$ for MAFC is described in Alg. 3. Procedure specific decision functions are required, with D_f^2 (Alg. 4) used for 2AFC and D_f^m (Alg. 1) used for MAFC. Each individual image evaluation is a trial. The value returned by ϕ_T represents one point on an item-response curve, which is the computed accuracy over all trials (one trial per class).

An item-response curve is the set of x, y coordinates that represent the model behavior for a set of stimuli. Each x, y value represents a perturbation level and accuracy of the model's performance. Note that traditional psychophysics experiments with live test subjects often apply a psychometric function to interpolate between the points to generate the curve. To approximate a psychometric function for better interpretability, we applied rectangular smoothing (*i.e.*, unweighted sliding-average smoothing) with a window size of 15 while padding the curve with repeated edge values.

The final step generates item-response curves using the function $\mathcal{C}_T(\phi, D, V, n, b_l, b_u)$. The procedure is simple, and only requires a repeated execution of ϕ_T for each stimulus level. Its steps are shown in Alg. 5. The procedure will create a set of stimulus levels starting with a lower bound, b_l, and ending with an upper bound b_u. b_l is the closest stimulus level to the preferred view and b_u is the stimulus level that is farthest away. The parameter n is the number of stimulus levels to use. Typically in visual pyschophysics, log-spaced stepping is used for finer-grained evaluation near the canonical view; the same strategy is used for preferred view.

IV. Experiments

The first goal of our experiments was to demonstrate PsyPhy as a large-scale psychophysics evaluation framework. To do this, we processed millions of procedurally rendered 3D scenes and 2D images that were perturbed. The second goal was to demonstrate the utility of procedural graphics for large-scale psychophysics experiments. Thus we broke our data up into two sets: natural scenes and rendered scenes. Our final goal was to evaluate the strengths and weaknesses of well-known CNN models. To do this, we looked at model behavior for 2AFC and MAFC tasks, the behavior of dropout at test time [43] under perturbations, and comparisons to

Algorithm 4 $D^2_f(i, p, q)$, best match decision of the final feature layer of a CNN. Used for 2AFC.

```
Input: $f$, a single pre-trained network model
Input: $i$, an input image
Input: $p$, the expected positive image
Input: $q$, the expected negative image
$W_i = f(i)$    ▷ gather activations from final feature layer
$W_p = f(p)$
$W_q = f(q)$
$\varsigma_p = r(W_i, W_p)$    ▷ Pearson's Correlation
$\varsigma_q = r(W_i, W_q)$
if $\varsigma_p > \varsigma_q$ then    ▷ if incorrect selection, negate response
    $\varsigma = \varsigma_p$
else
    $\varsigma = -1 * \varsigma_q$
end if
return $\varsigma$, the decision score
```

Algorithm 5 $\mathcal{C}_T(\phi, D, V, n, b_l, b_u)$: an item-response curve generation function for any type of decision function

```
Input: $\phi$, an item-response point generator
Input: $f$, an input model
Input: $V$, a vector of preferred views
Input: $n$, the number of stimulus levels
Input: $b_l$ and $b_u$, the lower and upper bound values of the
    stimulus levels
Let $S$ be $n$ log-spaced stimulus levels from $b_l$ to $b_u$
$I = \bigcup_{s \in S} \{\phi_T(D, V, s)\}$
return $I$, the item-response curve
```

human behavior. In all of our experiments, we chose to use five convolutional neural network models that were pre-trained on ImageNet 2012 [5]: AlexNet [6], CaffeNet [17], GoogleNet [18], VGG-16, and VGG-19 [19]. The complete set of plots and more details on the methods can be found in the supplemental material for this article.

Data Generation. For the natural scene experiments, we perturbed images from the ImageNet 2012 [5] training dataset, which consists of $\sim$1.2M million images and 1K classes. Using the training set instead of the testing set gives each model an intentional bias towards "expert" performance. The following transformations were applied: Gaussian blur, linear occlusion, salt & pepper noise, brightness, contrast, and sharpness. For each condition, we created 200 perturbed images starting with the preferred view and log-spaced stepped towards increasing difficulty. The result was 201 images per class per network, or 201K images per network, or $\sim$1M images per condition. In total, $\sim$9M images were evaluated.

For the experiments with rendered images, we selected 40 3D objects from the Blend Swap [44] library that corresponded to classes in ImageNet (see supp. material for a list of the classes). For each of the 3D objects, we randomly rendered 100K uniformly distributed x, y, z rotations and scales, resulting in 4M images. After each preferred view was selected from that set, the following transformations were applied by our graphics engine: rotations in the x, y, z dimensions, and scale. All were applied in the positive and negative direction. In addition, all of the transformations from the 2D natural image experiment were repeated using the rendered preferred views. For each of the 3D transformations, we rendered 200 images starting with the preferred view and log-spaced stepped towards increasing difficulty. The result was 201 images per class per network, or $\sim$8K images per network, or $\sim$40K images per transformation. The additional 2D transformations resulted in a total of $\sim$362K images, which brought the rendered image total to $\sim$683K evaluated images.

2AFC Experiments. The motivation for the 2AFC experiments is twofold: (1) to test decision making at a fundamental level via activation matching (*i.e.*, to not just look at class labels), and (2) to test a precise implementation of a well-known match-to-sample task. Given that this setting is just two instances of pair-matching, we had initially expected models to perform relatively well under a variety of perturbations. The experiments included the use of both natural scenes and rendered scenes as stimuli.

Model behavior was only very stable for the 3D rotation (Fig. 2 left and Supp. Fig. 1), and contrast and sharpness (Supp. Figs. 3 & 4) transformations. The rest of the transformations induced more erratic behavior, with accuracy declining below 80%. For example, Gaussian blur (Fig. 2 center) was very detrimental to model accuracy, even in this limited matching setting. This tells us something about the receptive fields of the convolutional layers of the networks: they are not large enough to tolerate even modest levels of blur. Also interesting were results for the 3D-specific scale perturbation that let us isolate specific failure modes. For example, when scale decreases to 20% of the original size, the object structure is still clearly visible, but accuracy drops to $\sim$60% or lower for all networks (Supp. Fig. 1). This is an observation that could not have been made by looking at summary statistics.

What about the differences in behavior across networks? Are they significant? When examining the item-response curves with the the 95% confidence interval plotted (Supp. Figs. 2, 5, 6 & 8) all network behavior consistently demonstrates the same trends for each transformation type across perturbations. While it is commonly believed that architectural innovation is advancing deep learning, this finding indicates that model behavior is more influenced by the training data, which was the same for all models in these experiments. For the VGG networks, this suggests that additional layers — beyond a certain point — do not imply better performance under degrading conditions. Likewise, switching the order of the pooling and normalization layers in CaffeNet and AlexNet [45] does not imply better performance under degrading conditions.

MAFC Experiments. The motivation for the MAFC experiments is to evaluate a task that is more closely aligned to the multi-class classification task the models were originally trained for. Given that there are $1,000$ choices in this setting instead of just two, we expected models to perform much worse under the transformations. And this is exactly what we observed in the results (Fig. 3 and Supp. Figs. 9, 11, 12 & 15). For instance, compare the plot for positive rotation in the x-axis (Fig. 3 bottom-left) to the corresponding plot in Fig. 2. For this transformation type, the networks that

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2018.2849989, IEEE Transactions on Pattern Analysis and Machine Intelligence

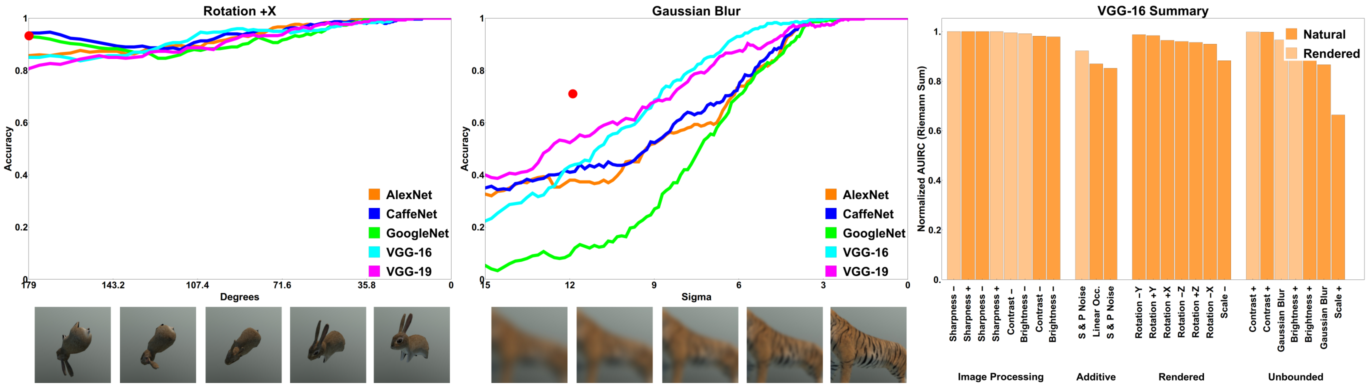

Fig. 2. (Left and Center) a selection of item-response curves for the 2AFC task. These rendered scene experiments reflect the accuracy across 40 classes. Each experiment used five well-known CNNs [6], [17]–[19]. A perfect curve would be a flat line at the top of the plot. The images at the bottom of each curve show how the perturbations increase from right to left, starting with no perturbation (*i.e.*, the original image) for all conditions. The red dot indicates mean human performance for a selected stimulus level. (Right) a summary plot of all VGG-16 2AFC item-response curves using AUIRC (Riemann sum) as a summary statistic, normalized by the total area above and below the curve. These plots (as well as the next sets in Figs. 3 & 4) are best viewed in color.

only experienced moderate dips in performance for the most extreme perturbations in the 2AFC case fall to under $\sim$20% accuracy at points. A caveat to the MAFC decision function is that because it is patterned after the classification task in computer vision, it only uses class labels to make its decisions. Thus it leaves out the layer-specific activation information that was used in the 2AFC case. This highlights an important trade-off that can occur when designing decision functions for psychophysics evaluations: task fidelity versus task difficulty.

Curiously, there are large asymmetries for some of the transformations with increasing and decreasing perturbation levels. See the plots for brightness, contrast, and sharpness (Fig. 3 top-left and top-center, and Supp. Figs. 11 & 12). Contrast is a particularly intriguing case. As a transformation, contrast is a non-linear single pixel-level operation applied globally to an image. In the positive direction, contrast is increased, and the performance of each network degrades rapidly (Fig. 3 top-center). In the negative direction, contrast is decreased, but the performance of each network remains relatively stable until the objects have very low contrast (Fig. 3 top-left). This suggests a contrast sensitivity problem under the MAFC decision function that is the opposite of what human patients with contrast-related vision deficits struggle with. There is a positive aspect to this finding — while diminished contrast sensitivity may induce night-blindness in a human driver, CNN-based autonomous driving systems can be expected to operate more effectively in the dark.

Cross-Perturbation Comparison. To facilitate comparison across perturbations, we generated one summary plot for each set of 2AFC (Fig. 2 right) and MAFC (Fig. 3 bottom-right) experiments. Each plot is generated using an area under the item-response curve (AUIRC) summary statistic, calculated with a midpoint Riemann sum and then normalized to unit space. This is similar in spirit to area under the curve in an ROC setting. A bar representing perfect performance has $y = 1.0$. A benefit of using AUIRC allows comparisons across perturbations without making assumptions about the underlying shape of the item-response curve. While model performance can effectively be compared using AUIRC, caution should be taken when comparing unbounded parameters (*e.g.*, σ for Gaussian blur) as such a comparison is dependent on the selected bound the experimenter has chosen.

Dropout Experiments. The experiments we have looked at thus far assume deterministic outputs. What about settings with stochastic outputs that support uncertainty in decision making? Gal and Ghahramani [43] introduced dropout *at testing time* to implement Bayesian inference in neural networks. What sort of variability does this lead to under various transformations and perturbation levels, and what does this tell us about the certainty of the experiments above? The setup for these experiments is identical to the setup for the MAFC experiments (including preferred views) except that during evaluation we applied dropout at test time to the Caffe version of AlexNet (which was also trained with dropout). Deploying the pre-trained model for each test, we dropped out 50% (because the model is large) of the neurons in layers *fe6* and *fe7* by uniformly randomly setting their activations to zero. This is repeated with 5 different random seeds for each transformation except salt & pepper noise and linear occlusion, which were not performed due to randomness in their underlying perturbation functions.

As anticipated, some variability in the base model performance was introduced (Fig. 4 and Supp. Figs. 17-20). But importantly, most runs still demonstrated a large measure of consistency (*e.g.*, Fig. 4) across the ranges of perturbations, indicating higher degrees of model certainty. This is a good stability property — when a model fails, it does so in an expected way across different dropout configurations, lending credibility to the initial characterizations of the behavior in the earlier experiments. More variability was observed for the rendered objects versus the natural scenes. This can be attributed to the use of the 3D objects that were outside of the training set for all of the models. The maximum difference observed between points from two runs for any transformation was 16.5% for sharpness applied to 3D objects (Supp. Fig. 19). In over half the cases, the maximum difference was over 10%.

Human Comparisons. Using psychophysics, model performance can be directly compared to human performance. To obtain human data points (red dots) for Figs. 2-4, we conducted a study with 24 participants (21 for contrast). Each participant

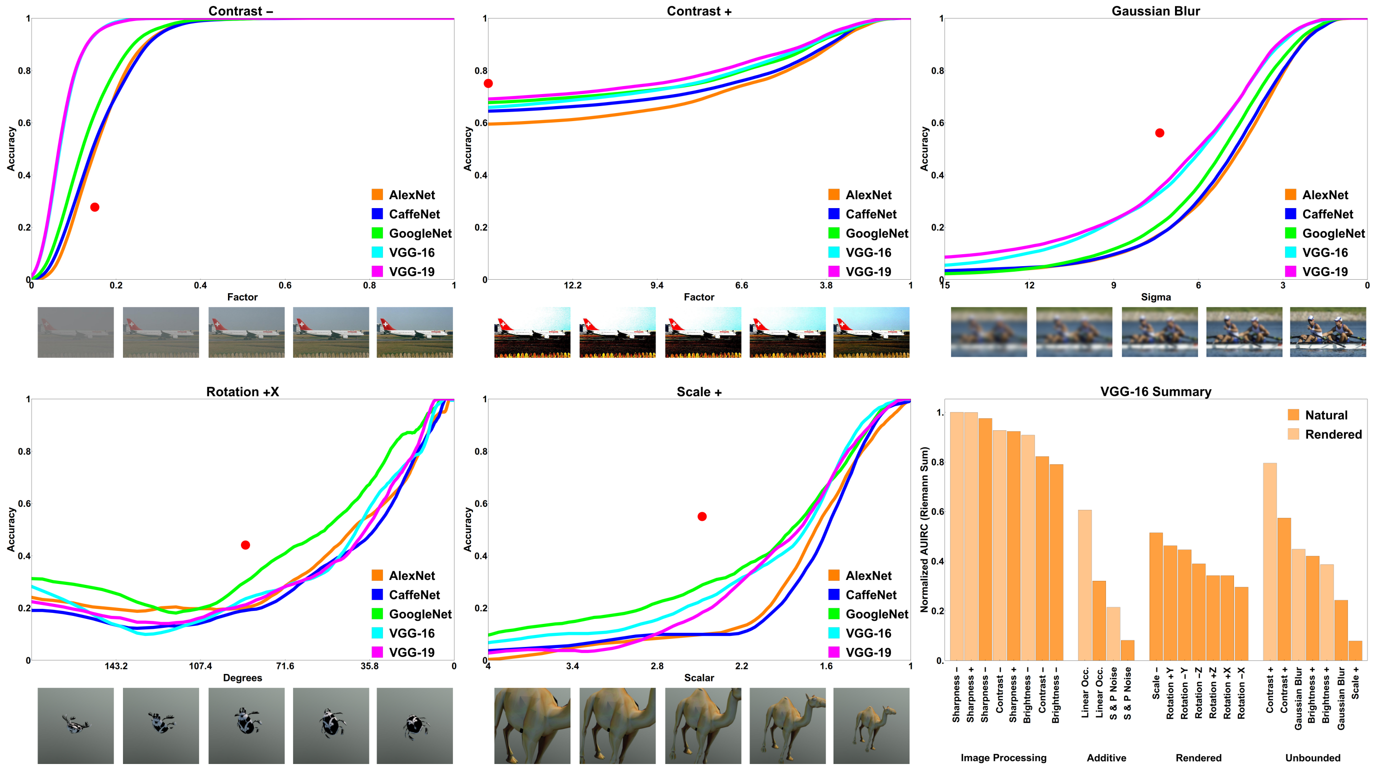

Fig. 3. A selection of item-response curves for the MAFC task. (Top) natural scenes. (Bottom-Left and Center) rendered scenes. The top-left can be directly compared to the top-center as well as the bottom-left to its corresponding plot in Fig. 2. (Bottom-Right) MAFC summary plot for VGG-16.

performed the 2AFC task as described above, but to mitigate fatigue, only 5 trials with a fixed psychophysical parameter setting were given for each of the transformations from Fig. 2. The participants also performed the MAFC task, but were limited to 3 choices instead of the full 1000 classes to make the task tractable. For those experiments, participants performed in 5 trials with a fixed psychophysical parameter setting for each transformation in Fig. 3. The original images for each trial were chosen randomly from the VGG-16 preferred views such that each class was only used one time for each participant in order to prevent participants from learning while performing the task. For all trials on both tasks, the sample images were presented for 50ms and subjects had unlimited time to answer.

Even without generating a full psychometric curve for the human subjects, it was apparent that only two out of eleven experiments showed any relative consistency between human and model performance (Fig. 2 left and brightness increasing in Supp. Fig. 1). While human performance was superior to model performance in most cases, there were two cases where humans were worse than the models: decreasing contrast (Fig. 3 left; for analysis, see MAFC experiments) and increasing brightness (Supp. Fig. 3). Brightness adjustment in image processing is a constant multiplicative change to all the pixel values, which preserves edges and allows the networks to recognize the geometry of an object almost until saturation is reached. Humans were also good at this task, but were still ∼9% worse than VGG-19 for the perturbation level analyzed.

V. Discussion

In visual psychophysics, we have a convenient and practical alternative to traditional dataset evaluation. However, the use of psychophysics testing and datasets are not mutually exclusive. One needs datasets to form a training basis for any data-driven model. Moreover, there is major utility to having a large amount of such data — this is essential for making machine learning capture enough intraclass variance to generalize well to unseen class instances. Data augmentation [6], [46] is an obvious strategy for leveraging the rendered images that were problematic for a model during psychophysics testing to expand the scope of the training set. However, this has diminishing returns as datasets grow to sizes that exceed available memory (or even disk space) during training. Using more limited training data and reinforcement learning that optimizes over item-response curves to correct for recognition errors is likely a better path forward.

Recent research has shown that CNNs are able to predict neural responses in the visual cortex of primates [32]. This, coupled with excellent benchmark dataset results across multiple recognition domains, suggests that good progress is being made towards reaching human-like performance. As a strong counterpoint, our psychophysics experiments show that the current most popular CNN models sometimes fail to correctly classify images that humans do not make mistakes on. What is missing from the models that is causing this behavioral discrepancy? With psychophysics as a guide, we can more easily discover what is missing — making it harder for us to be fooled by the person inside of the Chinese room.

Acknowledgment

The authors thank Lucas Parzianello for helping import Blendswap models into PsyPhy and Brian Turnquist for providing feedback on an early draft of this work. Funding was

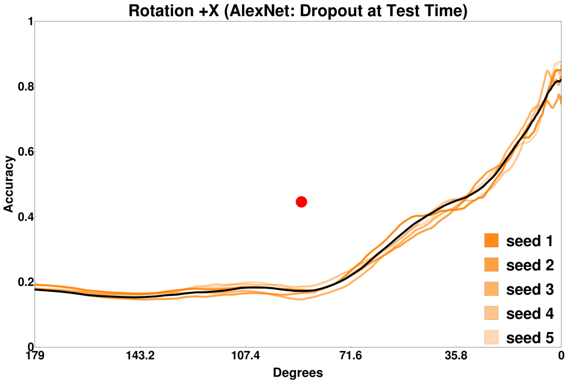

Fig. 4. Item-response curves for five different runs of an AlexNet model with dropout applied at test time [43] for a 3D rotation transformation. The black line indicates the mean of the five AlexNet curves. The maximum difference between points on any two curves in this plot is 12.2%.

provided under IARPA contract #D16PC00002, NSF DGE #1313583, and NSF SBIR Award #IIP-1621689. Hardware support was generously provided by the NVIDIA Corporation.

References

[1] Z.-L. Lu and B. Dosher, *Visual Psychophysics: From Laboratory to Theory*. MIT Press, 2013.

[2] F. Kingdom and N. Prins, *Psychophysics: a Practical Introduction*. Academic Press, 2016.

[3] J. R. Searle, "Minds, brains, and programs," *Behavioral and Brain Sciences*, vol. 3, no. 03, pp. 417–424, 1980.

[4] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla *et al.*, "ImageNet large scale visual recognition challenge," *IJCV*, vol. 115, no. 3, pp. 211–252, 2015.

[5] ImagetNet Large Scale Visual Recognition Challenge 2012 (ILSVRC2012), "http://image-net.org/challenges/LSVRC/2012/index," Accessed: 2016-10-12.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *NIPS*, 2012.

[7] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *IEEE CVPR*, 2011.

[8] T. Tommasi, N. Patricia, B. Caputo, and T. Tuytelaars, "A deeper look at dataset bias," ser. LNCS, J. Gall, P. Gehler, and B. Leibe, Eds. Springer, 2015, vol. 9358, pp. 504–516.

[9] S. E. Embretson and S. P. Reise, *Item response theory for psychologists*. Lawrence Erlbaum Associates, Inc., 2000.

[10] S. Hecht, S. Shlaer, and M. H. Pirenne, "Energy, quanta, and vision," *The Journal of General Physiology*, vol. 25, no. 6, pp. 819–840, 1942.

[11] J. K. Bowmaker and H. J. Dartnall, "Visual pigments of rods and cones in a human retina," *The Journal of Physiology*, vol. 298, no. 1, pp. 501–511, 1980.

[12] B. Duchaine and K. Nakayama, "The Cambridge face memory test: Results for neurologically intact individuals and an investigation of its validity using inverted face stimuli and prosopagnosic participants," *Neuropsychologia*, vol. 44, no. 4, pp. 576–585, 2006.

[13] J. B. Tenenbaum, C. Kemp, T. L. Griffiths, and N. D. Goodman, "How to grow a mind: Statistics, structure, and abstraction," *Science*, vol. 331, no. 6022, pp. 1279–1285, 2011.

[14] I. Yildirim, T. D. Kulkarni, W. A. Freiwald, and J. B. Tenenbaum, "Efficient and robust analysis-by-synthesis in vision: A computational framework, behavioral tests, and modeling neuronal representations," in *Annual Conference of the Cognitive Science Society (CogSci)*, 2015.

[15] T. D. Kulkarni, P. Kohli, J. B. Tenenbaum, and V. Mansinghka, "Picture: A probabilistic programming language for scene perception," in *IEEE CVPR*, 2015.

[16] J. Wu, C. Zhang, T. Xue, W. T. Freeman, and J. B. Tenenbaum, "Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling," in *NIPS*, 2016.

[17] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," *arXiv preprint arXiv:1408.5093*, 2014.

[18] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *IEEE CVPR*, 2015.

[19] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[20] M. Riesenhuber and T. Poggio, "The individual is nothing, the class everything: Psychophysics and modeling of recognition in object classes," MIT, Tech. Rep. AIM-1682, October 2000.

[21] ——, "Hierarchical models of object recognition in cortex," *Nature Neuroscience*, vol. 2, no. 11, pp. 1019–1025, 1999.

[22] S. Eberhardt, J. Cader, and T. Serre, "How deep is the feature analysis underlying rapid visual categorization?" in *NIPS*, 2016.

[23] R. Geirhos, D. H. J. Janssen, H. H. Schütt, J. Rauber, M. Bethge, and F. A. Wichmann, "Comparing deep neural networks against humans: object recognition when the signal gets weaker," *arXiv preprint 1706.06969*, 2017.

[24] H. E. Gerhard, F. A. Wichmann, and M. Bethge, "How sensitive is the human visual system to the local statistics of natural images?" *PLoS Computational Biology*, vol. 9, no. 1, p. e1002873, 2013.

[25] W. J. Scheirer, S. E. Anthony, K. Nakayama, and D. D. Cox, "Perceptual annotation: Measuring human vision to improve computer vision," *IEEE T-PAMI*, vol. 36, no. 8, August 2014.

[26] L. Germine, K. Nakayama, B. C. Duchaine, C. F. Chabris, G. Chatterjee, and J. B. Wilmer, "Is the web as good as the lab? comparable performance from web and lab in cognitive/perceptual experiments," *Psychonomic Bulletin & Review*, vol. 19, no. 5, pp. 847–857, 2012.

[27] C. Vondrick, H. Pirsiavash, A. Oliva, and A. Torralba, "Learning visual biases from human imagination," in *NIPS*, 2015.

[28] A. J. O'Toole, P. J. Phillips, F. Jiang, J. Ayyad, N. Penard, and H. Abdi, "Face recognition algorithms surpass humans matching faces over changes in illumination," *IEEE T-PAMI*, vol. 29, no. 9, pp. 1642–1646, 2007.

[29] A. J. O'Toole, X. An, J. Dunlop, V. Natu, and P. J. Phillips, "Comparing face recognition algorithms to humans on challenging tasks," *ACM Transactions on Applied Perception (TAP)*, vol. 9, no. 4, p. 16, 2012.

[30] P. J. Phillips and A. J. O'Toole, "Comparison of human and computer performance across face recognition experiments," *Image and Vision Computing*, vol. 32, no. 1, pp. 74–85, 2014.

[31] C. F. Cadieu, H. Hong, D. Yamins, N. Pinto, N. J. Majaj, and J. J. DiCarlo, "The neural representation benchmark and its evaluation on brain and machine," in *ICLR*, 2013.

[32] D. Yamins, H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert, and J. J. DiCarlo, "Performance-optimized hierarchical models predict neural responses in higher visual cortex," *Proceedings of the National Academy of Sciences*, vol. 111, no. 23, pp. 8619–8624, 2014.

[33] H. Hong, D. Yamins, N. J. Majaj, and J. J. DiCarlo, "Explicit information for category-orthogonal object properties increases along the ventral stream," *Nature Neuroscience*, vol. 19, no. 4, pp. 613–622, 2016.

[34] R. T. Pramod and S. P. Arun, "Do computational models differ systematically from human object perception?" in *IEEE CVPR*, 2016.

[35] D. Hoiem, Y. Chodpathumwan, and Q. Dai, "Diagnosing error in object detectors," in *ECCV*, 2012.

[36] M. J. Wilber, V. Shmatikov, and S. Belongie, "Can we still avoid automatic face detection?" in *IEEE WACV*, 2016.

[37] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *ICLR*, 2014.

[38] A. Nguyen, J. Yosinski, and J. Clune, "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images," in *IEEE CVPR*, 2015.

[39] A. Bendale and T. E. Boult, "Towards open set deep networks," in *IEEE CVPR*, 2016.

[40] V. Blanz, M. J. Tarr, and H. H. Bülthoff, "What object attributes determine canonical views?" *Perception*, vol. 28, no. 5, pp. 575–599, 1999.

[41] J. W. Peirce, "PsychoPy: Psychophysics software in python," *Journal of Neuroscience Methods*, vol. 162, no. 1-2, pp. 8–13, 2007.

[42] W. Jakob, "Mitsuba https://pillow.readthedocs.io/en/3.0.0/index.html," Accessed: 2016-11-05.

[43] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *ICML*, 2016.

[44] Mitsuba renderer, "http://www.blendswap.com," 2010.

[45] BVLC Caffe, "https://github.com/BVLC/caffe," Accessed: 2016-11-10.

[46] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," in *BMVC*, 2014.