Short Paper

Measuring Human Perception to Improve Open Set Recognition

Jin Huang, *Student Member, IEEE*, Derek Prijatelj, *Student Member, IEEE*, Justin Dulay, *Student Member, IEEE*, and Walter Scheirer, *Senior Member, IEEE*

***Abstract*—The human ability to recognize when an object belongs or does not belong to a particular vision task outperforms all open set recognition algorithms. Human perception as measured by the methods and procedures of visual psychophysics from psychology provides an additional data stream for algorithms that need to manage novelty. For instance, measured reaction time from human subjects can offer insight as to whether a class sample is prone to be confused with a different class — known or novel. In this work, we designed and performed a large-scale behavioral experiment that collected over 200,000 human reaction time measurements associated with object recognition. The data collected indicated reaction time varies meaningfully across objects at the sample-level. We therefore designed a new psychophysical loss function that enforces consistency with human behavior in deep networks which exhibit variable reaction time for different images. As in biological vision, this approach allows us to achieve good open set recognition performance in regimes with limited labeled training data. Through experiments using data from ImageNet, significant improvement is observed when training Multi-Scale DenseNets with this new formulation: it significantly improved top-1 validation accuracy by 6.02%, top-1 test accuracy on known samples by 9.81%, and top-1 test accuracy on unknown samples by 33.18%. We compared our method to 10 open set recognition methods from the literature, which were all outperformed on multiple metrics.**

***Index Terms*—Computer vision, open set recognition, novelty detection, visual psychophysics, deep learning.**

I. Introduction

Open Set Recognition (OSR) is a task that is extremely challenging for computer vision algorithms [1], but something that can be effortlessly performed by humans without the need for very large labeled datasets [2]. In machine learning-based computer vision, OSR is defined as novelty detection coupled with closed set classification, where novelties are visual information not seen at training time that should not inhibit classification performance [3]. At the class level, images from classes that are seen in both the training and testing phases are called known classes, while those that are not seen in training and only encountered in testing are considered to be unknown classes (i.e., novel classes).

While some existing systems can detect unknown classes based on previous known knowledge [1], humans can not only recognize unknown instances but also avoid the under-generalization problem machine learning models face when fitting to known class data using just supervised labels [4]. Psychologists study this phenomenon by using the methods and procedures of visual psychophysics to measure the human behavior associated with a particular task [5]. Reaction time (RT) is one of the most diagnostic measurable behaviors because it reveals patterns of difficulty in the data (e.g., it is fast to recognize something that is familiar, while more ambiguous cases will take longer to be recognized).

In this paper, we consider the possibility of incorporating measurements of human reaction time into the OSR process to improve model performance. The subtleties of human perception measured by psychophysics methods have already made in-roads in other computer vision tasks [6], [7], [8], making this strategy an attractive target for OSR. The fact that humans can separate objects that belong to a task from objects that do not in a better way than machines indicates that incorporating psychophysical measurements of human reaction time into the model training process may provide more information to learn from, and thus better performance. The challenge is in designing a biologically-inspired neural network architecture with its own variable reaction time and conditioning it to match human behavior as closely as possible.

The first step to solving this problem is collecting data that are useful in the context of novelty management. Boult et al. [9] have recently argued that much of the existing OSR work consists of ill-defined novelty problems, making both data collection and algorithm design more difficult than is necessary. Thus they introduce a formal taxonomy of novelty and suggest that the perception of novelty can be decoupled from the "ground-truth" novelty value of objects in the environment. This means that something can be novel to an agent (artificial or human) that may not be truly novel in the environment (i.e., it has not been learned yet). In-line with this, we designed a human behavioral experiment which gauged the human perception of class-level information via reaction time measurement (Fig. 1, top panel). We collected over 200,000 psychophysical reaction time measurements using this experiment design for a partition of ImageNet containing 335 total classes [10]. Compared to prior work in OSR [11], [12], [13], [14], [15], this dataset is far more realistic and challenging than the open set partitions of MNIST, SVNH, CIFAR10 and other small datasets that are commonly found in the literature. But it is intentionally more limited in available training images compared to the common ImageNet-based OSR training regimes.

The second step is the formulation of the machine learning strategy. After obtaining knowledge of the correlation between reaction time and OSR performance for individual images, we designed a new

Manuscript received 6 September 2022; revised 12 February 2023; accepted 14 April 2023. Date of publication 27 April 2023; date of current version 4 August 2023. This research was sponsored in part by the National Science Foundation (NSF) under Grant CAREER-1942151 and in part by the Defense Advanced Research Projects Agency (DARPA) and the Army Research Office (ARO) under multiple contracts/agreements including under Grants HR001120C0055, W911NF-20-2-0005, W911NF-20-2-0004, HQ0034-19-D-0001, and W911NF2020009. Recommended for acceptance by R. P. Wildes. *(Corresponding author: Walter Scheirer.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the University of Notre Dame IRB under Application No. 18-01-4341.

The authors are with the Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN 46556 USA (e-mail: jhuang24@nd.edu; dprijate@nd.edu; jdulay@nd.edu; walter.scheirer@nd.edu).

This article has supplementary downloadable material available at https://doi.org/10.1109/TPAMI.2023.3270772, provided by the authors.

Digital Object Identifier 10.1109/TPAMI.2023.3270772

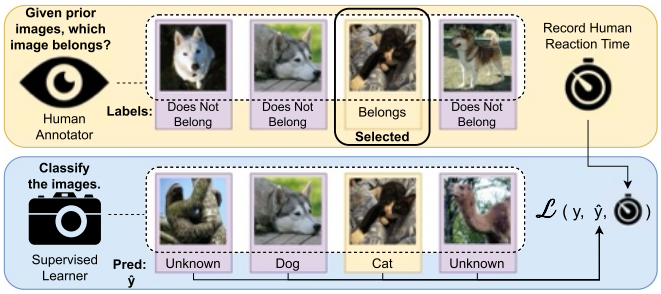

Fig. 1. In this work, we conduct behavioral experiments to gauge the human perception of objects that belong or don't belong to a task. Through the use of visual psychophysics, human reaction time is measured across tasks (top panel). These measurements are then used in a novel psychophysical loss function to train biologically-inspired deep networks with a variable reaction time property, leading to better novelty detection and multi-class classification performance in open set recognition settings (bottom panel).

psychophysical loss function that can enforce more consistent behavior between humans and machines when training a deep neural network, thus improving the performance on OSR problems (Fig. 1, bottom panel). To ensure the best match between the newly introduced loss function and the task, we chose to use the Multi-Scale DenseNet [16] (MSD-Net) architecture as a proof-of-concept. MSD-Net has five classifiers at different depths in its architecture, making it possible to measure model reaction time by observing the classifier that serves as the exit point for a particular image. We believe that coupling large-scale psychophysically annotated data with a loss function that makes use of behavioral measurements like reaction time to induce similar behavior in models is a viable path forward for the OSR problem.

In summary, the major contributions of this paper are: (1) A new dataset for the OSR problem containing reaction time measurements for a challenging partition of ImageNet. This data is analyzed to draw out patterns of difficulty not evident from the original class-level labels. This is the first large-scale human behavior study carried out to explore the use of psychophysical measurements for OSR, and it provides a wealth of data for other researchers working on this problem and related ones.[1] (2) A general strategy for utilizing measured human reaction time for training deep networks for OSR, implemented as a psychophysical loss function that can improve the performance of any neural network architecture that supports variable reaction times for different inputs. (3) A specific implementation of the loss strategy that takes advantage of the special structure of the MSD-Net architecture. (4) Extensive experimentation over the OSR partition of ImageNet that balances class diversity and limited training data, with comparison made to 10 recent OSR approaches from the literature. MSD-Net models that are trained with the proposed psychophysical loss function are shown to significantly outperform prior work.

II. Related Work

Open Set Recognition. The development of capabilities that can detect that a sample is novel is a classic problem within the broader field of pattern recognition [17]. But merely detecting novelty is a limited task — classification is far more useful mode of operation in machine learning. Most recognition work in computer vision has been conducted in a closed set classification mode, meaning all classes are known at both training and testing time. In contrast, OSR is a more realistic scenario, where partial knowledge of the world is present during training, and unknown phenomena are guaranteed to appear during testing. The difficulty has been in finding effective approaches for distinguishing between known and unknown samples [1].

Scheirer et al. initially formalized the OSR problem [3] and introduced the concept of *openness*, which characterizes the potential difficulty of open set problems. Importantly, they defined a notion of open space risk, based on the assumption that the farther away a sample is from the support of known training data in a feature space, the riskier it is to assign a known class label to it. Following this definition, a series of standalone classifiers was developed, including the 1-vs-Set Machine [3], P_I-SVM [18], and Weibull-Calibrated SVM (W-SVM) [19]. Moving beyond SVM-like classifers, Rudd et al. proposed the Extreme Value Machine (EVM) [12], which is an OSR classifier based on the statistical Extreme Value Theory (EVT) that supports incremental learning.

OpenMax [11] was the first OSR classifier that could be trained with a deep neural network, adapting EVT concepts to the activation patterns in the penultimate layer to perform OSR. Extending OpenMax, Generative OpenMax [20] trains a deep network with synthesized unknown data. Similarly, Counterfactual Images for Open Set Recognition (OSRCI) [21] uses a Generative Adversarial Network (GAN) to generate unknown images that are close to the training images but technically do not belong to any training class. Class Reconstruction Learning for Open Set Recognition (CROSR) [14] utilizes latent representations for reconstruction to improve performance on unknown samples. The Class Conditioned Auto-Encoder for Open Set Recognition (C2AE) [22] divides the training procedure into two sub-tasks, closed set classification and open set identification, and uses EVT to find a threshold for filtering unknown samples. Class Anchor Clustering for Open Set Recognition (CAC-OSR) [15] uses a distance-based loss function that forces known classes to form tight clusters around anchored class-dependent centers. Multi-Task OSR [23] combines a classifier and a decoder network with a shared feature extractor network within a multi-task learning framework for OSR.

Outlier Exposure (OE) is another strategy for OSR, where outlier data is provided to the model during the training process so it can generalize and ideally detect unknown samples. It is possible to leverage outlier samples to improve anomaly detection by training anomaly detectors against an auxiliary dataset of outliers [24]. Dhamija et al. [25] introduced a new evaluation metric that focuses on comparing the performance of multiple approaches in scenarios where unknowns are presented and proposed novel losses that are designed to maximize entropy for unknown inputs. Kong and Ramanan [26] introduced OpenGAN, which augments training outliers with fake unknown data synthesized by a generator trained to fool the discriminator.

Considering all of the above approaches, the human capacity for managing novelty has been missing. The previous work has focused on formalizing the OSR problem and developing methods that can improve OSR performance. However, none are informed by the measurement of human perception. This is surprising, considering that humans have a much stronger ability to perceive what belongs or doesn't belong to a task than computer algorithms. Different from previous OSR work, we suggest the use of visual psychophysics as a path forward.

Visual Psychophysics for Computer Vision. Some researchers have attempted to draw a connection between the human visual system and artificial neural networks, looking into the differences between various visual systems [27], as well as comparing the consistency between human vision and deep networks [28]. Furthermore, a growing amount of work has been studying how human behavior and perception can be related to solving computer vision problems. For example, DiCarlo et al. [29] highlighted how the human brain can perform object

[1]Data for this paper can be downloaded from: https://tinyurl.com/mrb577su. Code can be downloaded from: https://github.com/KiyoshiKAWASAKI/pami_osr.

recognition despite a difficult environment and many variances in the setting. After reviewing the evidence from human behavior to neural recordings, they proposed that neuronal and psychophysical data is necessary for better computational models. Medathati et al. [30] presented an overview of computational approaches to biological vision, and showed how new computer vision methods can be developed from biological insights. And studies have found that the human visual system takes different amounts of time to respond to stimuli of varied difficulty [31], [32]. All of these works touched on the notion that psychophysical methods show promise for computer vision algorithm development, but from the perspective of biology.

From the perspective of computer vision, Scheirer et al. proposed the idea of *Perceptual Annotation* [6] to make use of psychophysical measurements collected via crowdsourced means to train more accurate SVM classifiers. McCurrie et al. [7] collected behavioral data reflecting human judgments of subjective facial attributes in order to model them. Zhang et al. [33] suggested the use of human gaze measurements, another behavioral measurement type, for improving performance in object-related tasks. RichardWebster et al. [34], [35] proposed an evaluation framework for visual recognition models that constructs item-response curves made up of individual stimulus responses to find perceptual thresholds.

Most closely related to the work in this paper is that of Grieggs et al. [8], which introduces a psychophysical loss formulation for training artificial neural networks. In that work, behavioral experiments were conducted to collect reaction time data associated with the ability of reading in order to improve handwritten character recognition in historical documents. There are a few major differences between our work and [8]: (1) We develop a psychophysical loss function that incorporates several data streams and error measures. (2) Our work focuses on the general area of image classification instead of the relatively niche area of historical document processing. Not much expertise is required for our task, thus it is possible for us to conduct larger-scale human data collections. (3) Our goal is to utilize the psychophysical loss to improve a deep network's performance on OSR, instead of just achieving better closed set classification performance.

III. Psychophysical Study of Known Data

To better understand how humans perceive familiar images in order to be tolerant to novelty, we designed and conducted a study using Amazon Mechanical Turk to collect human reaction time (RT) data.[2] As the source image data, we used a subset of ImageNet [36] that contains 335 classes (called ImageNet335 in this paper) [10], which was created by the University of Maryland, Carnegie Melon University, and Columbia University as part of the DARPA Science of Artificial Intelligence and Learning for Open-world Novelty (SAIL-ON) program. For the purpose of novelty detection, we partitioned that dataset into two categories for machine learning. (1) *Known Classes:* the classes with distinctly labeled positive training examples that appear at both training and testing time; these are the non-novel classes. (2) *Unknown Classes:* the classes that are not labeled and are unseen in training; these are the novel classes used in testing.

With respect to the specific breakdown, we randomly selected 42 classes as unknown classes from ImageNet335, and the remaining 293 classes were used as known classes. In this study, we collected reaction time measurements for the detection of specific known samples in the midst of images from other classes. Knowing from previous work that a mix of original and psychophysically annotated training data tends to yield good results [6], [8], we randomly chose 40 classes from the known classes to use in the human behavior experiments.

[2] Approved by University of Notre Dame IRB under protocol 18-01-4341.

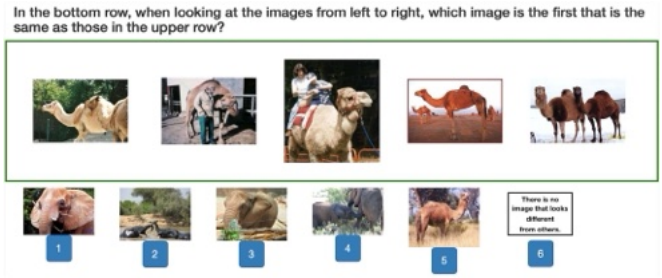

Fig. 2. A sample of one of the survey questions for collecting human reaction time for known images. The subjects are asked to look at both rows of images and select the first image in the bottom row that is of the same object class as the reference class in the top row. It is possible that no image from the reference class is shown, hence the sixth option in the bottom row. In this specific example, the subjects should select the fifth image.

Study Design. While *novelty* can be defined in different ways depending on the task and context, the problem we are trying to address in this research is class-level novelty. We designed a study that is suitable for collecting human reaction time measurements associated with known classes; we do this because unknown classes should not appear at training time in any capacity for a fair evaluation. Accordingly, we need a task that can provide information about the patterns of difficulty associated with known training samples as they interact with other classes. Here we can use other known class data as stand-in material for novel samples as RT measurements, which reflect difficulty, are made.

Each Mechanical Turk survey contained 25 questions, and subjects received 25 cents for completing a full survey. Fig. 2 shows an example of one survey question (the full survey process is detailed in Supp. Mat. Sec. 1.1), which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2023.3270772. In each question, two rows of images are shown to a subject. In the top row, five images are shown within a green box and are treated as reference images — all of them from the same class. In the bottom row, five additional images are shown. The task is for the subject to look at the images in the bottom row sequentially from left to right in order to find the first image that belongs to the reference image class. If the subject believes that such an instance isn't present in the bottom row, then they can indicate that as their response. A timer is started when a question populates the page and is stopped when the subject makes a decision. This is the recorded reaction time that is associated with the image that belongs to the reference class in the question. If the correct answer is the sixth option (a quality control measure), then the subject's recorded answer and RT will not be used for machine learning training.

We consider the reference images in the first row of a question as training data for a human, and the images in the second row as testing data. Training images for our machine learning task are drawn from the correct known class images from the second rows across questions and surveys. The intuition behind this design is that RT on known samples provides more information on what is non-novel beyond a class label, and if a model has a better understanding of what is known, it should at the same time be able to better judge what could be unknown by increasing the separation between these two categories. This is consistent with a recent observation that novelty detection in deep networks works by detecting the absence of familiar features as opposed to the presence of novelty [37]. Taking this idea a step further, the data from our study can be utilized effectively given a deep network

with a notion of variable reaction time for each input [16], which is trained to enforce behavioral fidelity with humans.

We pick one known class for the first row training data and another known class as a stand-in novel class that is different from the training data for the second row testing data. To be specific, in a question whose answer is not the sixth option, there are 5 images that are from a single class in the first row; in the second row, there are 4 images that are from a different class and 1 image that is from the same class as the first row. Such a design is straightforward for subjects to understand because there are only 2 classes shown at a time in a question, thus we expect to minimize noise in the results. Assuming we pick n classes for the surveys, each class is paired with itself and the rest of the $n-1$ classes. Using each pair, 20 questions are created, and based on the above there are n^2 different sets of surveys for the n classes chosen. In the experiments conducted there are 40 known classes, thus there are 1,600 sets of different surveys in total.

To filter out bad submissions, we also include five control questions in every survey, which means each survey has 25 questions in total. These control questions are designed to be trivially easy so that any diligent worker will most likely get them correct. We allow the subjects to answer at most two control questions incorrectly. Answers provided by unreliable subjects are filtered out and are not considered for the machine learning experiments. Subject were not able to answer the same question twice. Valid response times had to be under 28 seconds (the maximum RT after removing the largest RT values reflecting 5% of the total data) on the assumption that anything longer reflected an inattentive subject. Although we instructed the subjects to look at the images from left to right sequentially, there was no control to enforce this on their side. Thus we require that 5 subjects complete each question so that noise can be smoothed out by considering average reaction times for each image. To ensure the order of the images does not affect the reaction time entries, all the images are selected randomly when the survey is generated and the images do not follow any pre-set order.

Summary of Collected Data. We collected 211,074 RT measurements in total. After data cleaning and data pre-processing, we sampled 121,073 instance-level RTs for model training. We grouped the RTs that belong to the same image and calculated the average of them, and we used this average value to represent the RT for an image. Overall, there are 33,548 training samples from the 40 known classes, and 12,428 images among them have corresponding RTs.

After collecting the RT data, an important question for the subsequent machine learning work was whether to use class-level or sample-level behavioral information in the loss formulation. We started by looking at class-level RT by generating box plots summarizing the measurements by class pairings for each class (see Supp. Mat. Sec. 1.2), available online. To be useful for a loss function, large relative differences between a class and its different pairings would need to be present. We observed that the variance in RT within a class across its various other-class pairings was small, and that the various pairings have similar minimum and median RT statistics. This indicates that there isn't enough RT information at the class-level to be useful for a loss function, as patterns of difficulty across different classes cannot be ascertained.

Next, we turned to the sample-level behavioral information. We plotted the distribution of RT measurements for these known samples, shown in Fig. 3. Looking at the distribution, there is a large amount of variance across samples from all of the classes. This means that some samples have longer RT measurements while others have shorter ones, regardless of the classes they come from. Therefore the understanding and usage of RT has to be taken down to the sample-level for the design of the loss function. Our RT data indicate that the human visual system is perceiving each sample differently. Using this finding as an intuition for machine learning, we make the following assumption: human RT is a significant indicator of the latent difficulty of perceiving an object relative to objects from other classes, meaning these measurements will be useful or supervised training because the training set captures those relative comparisons.

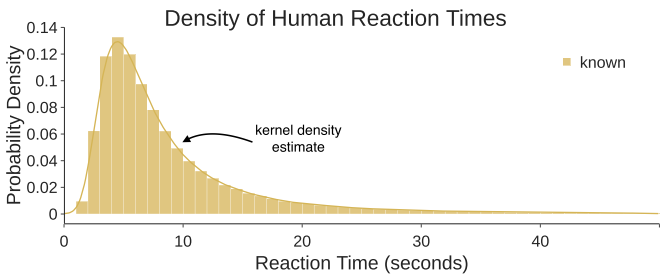

Fig. 3. This histogram and kernel density estimate shows the distribution of human reaction time for the data collected for known samples. The X-axis shows the range of reaction times after thresholding the long tail, which removed outliers. The Y-axis shows the probability of occurrence.

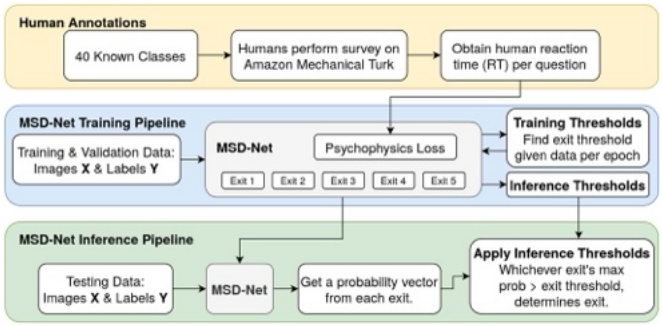

Fig. 4. Pipeline depicting the three components of the proposed machine learning process. The yellow box represents the human study for collecting RT. The blue box shows machine learning training. The green box illustrates the testing process. See Supp. Mat. Sec. 3.1 for additional detail on these, available online.

IV. Training a Model With Variable Reaction Time

With a large number of reaction time measurements associated with individual training points, we need some efficient way to make use of that data for training a deep network. The most straightforward approach has been to incorporate such data into a loss function [6], [8]. Unexplored before in the OSR literature, however, is a tighter coupling between human and model behavior when it comes to the reaction time associated with inputs. Thus we propose to enforce human behavioral fidelity in a model that supports dynamic reaction time during runtime (Fig. 4). This is accomplished via a new psychophysical loss function that encourages the model to mimic per instance human RT.

Multi-Scale DenseNet. Most network architectures use approximately the same amount of time to process all input images. Typically, networks only have one classifier at the end of the architecture, and it is extremely difficult, if not impossible, to discern any meaningful difference in model reaction time between samples. To be able to support different model reaction times and to train the model with human RT data, we chose to use the Multi-Scale DenseNet (MSD-Net) [16] architecture with a custom loss function. We also describe an experiment for a variant of ResNet-18 in Supp. Mat. Sec. 3.2, available online.

Different from many other networks that only have one classifier at the end, MSD-Net has five classifiers in its architecture. Utilizing these classifiers as model exits, we modify the network to output the

predictions after each classifier, and use a thresholding strategy to determine whether the network is confident enough to decide which class a sample should be classified into, or whether a sample is unknown. We start training by setting all 5 thresholds corresponding to each exit to zero. To update these thresholds, we extract the prediction scores on validation data every 5 epochs at each exit, taking the median of all the prediction scores for a specific exit as its threshold. These updates are necessary because the model behavior changes over time, and probabilities become higher as the model better describes the data.

We call these thresholds *training thresholds*. If the network produces a prediction score that is above the threshold and outputs a correct prediction, a sample is considered to be exiting the network. We consider it important that a sample has to satisfy both conditions during training because we need to enforce the constraint that the network confidently makes correct predictions. Training thresholds are updated based on the probability vectors from the exit loss described below.

Psychophysical Loss Formulation. The base loss used for training MSD-Net is the common multi-class classification cross-entropy loss. The cross-entropy loss is the function

$$\mathscr{L}_C(p,q) = -\sum_{y\in Y} p(y)\log q(y), \tag{1}$$

where p is the labeled probability (one-hot vector) of the class y from the closed set of classes Y and q is the MSD-Net predicted probability of the class y. Here we add to the cross-entropy loss by incorporating RT measurements via two additional psychophysical loss functions. This is accomplished via a weighted summation of the cross-entropy loss and a psychophysical loss.

The psychophysical loss is an exit loss

$$\mathscr{L}_E\left(x,\hat{y}\right) = \left|\mathscr{E}_{\text{target}(x)} - \mathscr{E}_{\hat{y}}\right|, \tag{2}$$

where $\mathscr{E}_{\hat{y}}$ is the exit integer index of the MSD-Net prediction $\hat{y}$, and $\mathscr{E}_{\text{target}(x)}$ is a lookup function that returns the target exit index for the sample x. The exit loss is designed to push the MSD-Net's reaction time for the sample to be proportional to the sample's target exit based on a binning strategy informed by the human RT measurements. To obtain the exit loss, the following two factors must be measured. (1) Expected exit: $\mathscr{E}_{\text{target}(x)}$. This value is assigned by measuring a discrete distribution of human reaction time. After removing the outliers from the human RT measurements and obtaining the minimum and maximum values, the entire range is discretized into five bins in accordance with the number of classifiers in the MSD-Net architecture. The cut-off thresholds defining the ranges covered by each bin are the quintiles of human RT in Fig. 3 and are shown in Supp. Table 2, available online. When a sample enters the network for training, the human RT associated with that sample is checked, and the corresponding exit index from the upper bound of the range is used as its $\mathscr{E}_{\text{target}(x)}$. For instance, if a known sample has an average human RT of 5.5 seconds, its $\mathscr{E}_{\text{target}(x)}$ will be 2. (2) Predicted exit for a sample determined by when it leaves the network: $\mathscr{E}_{\hat{y}}$. As a sample is processed through each classifier, the model checks the prediction scores as well as predicted label for that classifier. If the maximum prediction score produced is larger than the set threshold for an exit and the prediction is correct, the sample is considered to be leaving the network and the index of the exit is $\mathscr{E}_{\hat{y}}$.

The complete proposed psychophysical loss combines the cross entropy loss and the exit loss as a weighted summation

$$\begin{aligned}\mathscr{L}_\Omega\left(p,q,x,\hat{y},\vec{\omega}\right) = {} & \omega_C\mathscr{L}_C(p,q) \\ & + \omega_E\mathscr{L}_E\left(x,\hat{y}\right),\end{aligned} \tag{3}$$

where $\vec{\omega} = [\omega_C, \omega_E] : \omega_i \in \mathbb{R}$ is a vector of weights that correspond to the hyperparameter weighting of each component.

Training and Validation Pipeline. We take an original training dataset of images that do not have associated RTs and split it into training and validation sets with a ratio of 70% for training and 30% for validation. Available samples that do have associated measured RT are proportionally split via the same ratio into training and validation sets and merged with the samples that do not have any associated RT. We utilize an existing implementation of MSD-Net (https://github.com/kalviny/MSDNet-PyTorch), and send training and validation data into the network to train it with the psychophysical loss. Optionally, the weights in $\vec{\omega}$ can be set via domain knowledge or hyperparameter optimization. In our experiments, all the weights were set to 1.0 to assess base performance.

Testing Pipeline. After obtaining a model from training, we perform post-processing for novelty detection. First, we need to determine the threshold for each exit. This leads to a different set of thresholds than the training thresholds; we call these *inference thresholds*. To obtain this set of thresholds, the *best model* needs to be identified first. During the training phase, we save a model as well as the training and validation accuracy associated with it at every epoch. After training is done, the model that produced the highest top-1 validation accuracy is selected as the best model. We then run validation data through this best model and calculate the median of all the predicted scores for each exit respectively; these median scores are considered to be the inference thresholds. They remain static from this point forward because the best model is finalized. We then run all test samples from both known and unknown classes through the selected model, saving the prediction scores from each exit. Lastly, the inference thresholds are applied to these scores to find out when the samples leave, with corresponding exit indices recorded.

The exit strategies are slightly different for known samples and unknown samples. For all the known samples, a sample correctly exits the network when the maximum score is larger than a given inference threshold, and when the network makes a correct prediction. If a sample does not exit from one of the first 4 exits due to being under threshold each time, it comes to the final exit, where there are three possibilities for it. *(K1)* The maximum score of a sample is larger than the given threshold, and the prediction is correct: the sample exits and is correctly classified as a known class. *(K2)* The maximum score of a sample is larger than the given threshold, but the prediction is wrong: the sample exits and is classified as known, but associated with an incorrect class. *(K3)* The maximum score of a sample is smaller than the given threshold: the sample exits and is classified as unknown, making it a false negative.

As for testing unknown samples, there are only 2 possible cases for each sample. *(U1)* The maximum score is smaller than the given threshold at every exit: the sample is classified correctly as unknown. *(U2)* The maximum score at a particular exit is larger than the threshold given by that exit: the sample is wrongly classified as known, which makes it a false positive.

For testing known and unknown samples, we use the probability scores produced by the SoftMax function, and the exits solely decide whether a sample is known or unknown. During the testing phase, all samples are processed through the 5 exits and the probability scores from each exit are saved. The scores are then post-processed to obtain the final classification results.

V. Experiments and Results

Data and Evaluation Metrics. The training data breakdown can be found in Section III. For validation, 9,058 images are available from the 293 known classes. The test set consists of known and unknown images. The known test partition consists of the 293 classes seen in

TABLE I
RESULTS FOR THREE VARIATIONS OF THE LOSS USED TO TRAIN MSD-NET; SCORES ARE ACCURACY (%). FOR EACH LOSS, WE RUN EXPERIMENTS 5 TIMES USING 5 DIFFERENT SEEDS FOR TRAINING; THE SCORES SHOWN FOR EACH EXPERIMENT ARE THE AVERAGE OF 5 RUNS WITH STANDARD ERROR. $\mathscr{L}_P$ IS A PERFORMANCE LOSS [8] DESCRIBED IN SEC. 2 OF THE SUPP. MAT., AVAILABLE ONLINE ADDITIONAL RESULTS, INCLUDING RESULTS FOR A MODEL TRAINED WITH ALL 3 LOSSES AND COMPLETE RESULTS FOR ALL 5 RUNS, ARE IN SEC. 3 OF THE SUPP. MAT., AVAILABLE ONLINE

Split	Accuracy	$\mathscr{L}_C$	$\mathscr{L}_C + \mathscr{L}_P$	$\mathscr{L}_C + \mathscr{L}_E$
train	known top-1	**72.56**±0.47	**72.56** ± 0.27	72.10±0.42
train	known top-3	85.01±0.28	**85.09** ± 0.20	84.35±0.27
train	known top-5	89.02±0.22	**89.19** ± 0.15	88.37±0.22
valid.	known top-1	53.98±0.30	52.90 ± 0.51	**60.00**±0.27
valid.	known top-3	68.76±0.32	68.36 ± 0.20	**72.12**±0.28
valid.	known top-5	74.67±0.38	74.18 ± 0.13	**77.34**±0.18
test	known top-1	26.67±7.76	28.68 ± 6.91	**36.48**±0.43
test	known top-3	36.40±10.99	38.90 ± 9.87	**46.70**±0.73
test	known top-5	41.08±12.24	43.76 ± 11.14	**51.00**±0.89
test	unk. top-1	21.05±0.63	23.95 ± 2.98	**54.23**±0.48

TABLE II
TESTING RESULTS FOR THE PROPOSED LOSS AND OTHER BASELINES. BOLD NUMBERS INDICATE BEST PERFORMANCE FOR A METRIC. RESULTS IN RED HIGHLIGHT SPURIOUS BEST RESULTS DUE TO A CLASSIFIER BIASED TOWARDS DETERMINING SAMPLES ARE NOVEL

Algorithm	Test Unknown Acc.	Test Known Top-1 Acc
SVM [38]	55.04% ± 0.76	0.60% ± 0.07
P_I-SVM [18]	8.31% ± 0.12	0.28% ± 0.02
W-SVM [19]	9.21% ± 0.15	0.09% ± 0.01
EVM [12]	**99.39%** ± 0.00	0.45% ± 0.00
OpenMax [11]	32.41% ± 0.04	0.39% ± 0.00
OSRCI [21]	1.11% ± 0.01	0.36% ± 0.00
CROSR [14]	85.78% ± 0.00	1.92% ± 0.00
CAC-OSR [15]	94.55% ± 0.00	0.05% ± 0.00
MSD-Net $\mathscr{L}_C$	21.05% ± 0.63	26.67% ±7.76
MSD-Net $\mathscr{L}_C + \mathscr{L}_P$ [8]	23.95% ± 2.98	28.68% ± 6.91
MSD-Net $\mathscr{L}_C + \mathscr{L}_E$ (Eq. 3)	54.23% ± 0.48	**36.48%** ± 0.43

training providing a total of 336,453 new images. The unknown test partition consists of 42 classes unseen at training — all are considered to be part of one unknown class with 48,067 images.

Although various OSR works [14], [15], [21] use AUROC as a metric to evaluate performance, we believe that it is somewhat flawed for this problem. Importantly, it obscures the performance achieved through the use of a single threshold by combining a large range of thresholds to produce one score. Moreover, it does not provide a recommendation for selecting the best threshold. AUROC can be used fairly as a metric to assess a model's performance during training or validation, but there should only be one threshold per classifier used in the testing phase, as would be done in operation. Knowing this, we chose to use fixed thresholds obtained via experiments with the validation data as our inference thresholds.

As OSR is a task built on top of novelty detection by adding multi-class classification for known samples, both the detection and classification components can and should be evaluated separately. Thus we include metrics for both when reporting our results. For novelty detection, we consider accuracy when testing unknown samples. For multi-class classification, we use accuracy score and consider top-n accuracy for testing known samples. We report counts for True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN), as well as F1 Score and the Matthews Correlation Coefficient (MCC) in the Supp. Mat. Sec. 3 for comparison, available online.

Results: Contribution of Each Loss Function. Since the proposed psychophysical loss is composed of an exit loss added on top of a cross-entropy loss, we trained two models separately: one model is trained just with cross-entropy loss and the other is trained according to Eq. (3) (the proposed loss). To compare the performance of our loss formulation with the previous published psychophysics loss [8], we also trained a model with this *performance loss*. It is noted as $\mathscr{L}_P$, and is added on top of cross-entropy loss as another baseline.

We train each of the models using a single GPU with a batch size of 16 images shaped 224x224 and an initial learning rate of 0.1 for 200 epochs. As for the optimizer, we use stochastic gradient descent (SGD) with a momentum of 0.9 and a weight decay of 0.0001. All other parameters are set the same as the original implementation, and we did not heavily tune these training parameters. Evaluation takes place in a 5-fold manner, training models with 5 different fixed seeds in all cases. The average accuracy over all folds is considered as the final result.

Table I shows the training, validation and testing results for this experiment. All three configurations of the loss demonstrate roughly the same accuracy on known samples during training. When performance loss is added to cross-entropy loss, there is a noticeable uptick in test accuracy for both known and unknown samples. This indicates that measured RT is helping the trained model to generalize in the OSR setting. The increased performance in this case on known samples is consistent with prior work [8], but the increase in performance on unknown samples is a new finding. Better conditioning of the MSD-Net exit behavior yields even more performance improvement. For top-1 validation accuracy, the proposed loss of Eq. (3) outperforms the cross-entropy loss by 6.02%, and the combined cross-entropy and performance loss by 7.1%. Even more significant improvement is demonstrated in the testing phase. When testing on known classes, the proposed loss results in a top-1 accuracy of 36.48%, outperforming the two other loss configurations (26.67% and 28.68%). When detecting unknown samples, the proposed loss largely outperforms the others with 54.23% accuracy, compared to 21.05% and 23.95%. These results show that our proposed loss leads to models that are able to maintain good performance when testing with multiple known classes while simultaneously rejecting unknown samples.

Results: Other Baselines. We also compared our method with eight other OSR methods from the literature. With respect to standalone classifiers, we evaluated SVM [38] with thresholding, P_I-SVM [18], W-SVM [19], and the EVM [12] with MSD-Net features. With respect to deep learning-based methods, we evaluated OpenMax [11], OSRCI [21], CROSR [14], and CAC-OSR [15]. See Supp. Mat. Sec. 2 for descriptions of these approaches, available online. For these experiments, we used the same dataset used to evaluate MSD-Net training with different loss configurations. As we did not heavily tune hyper-parameters in our proposed approach to avoid overfitting, for a fair comparison we also refrained from doing so for all baseline methods. Table II shows results on known samples and unknown samples from the test set for all methods, including the different configurations of MSD-Net training.

While most of the methods only achieve performance that is only a little better than making a random choice among the 293 known classes, the different configurations of MSD-Net training achieve excellent top-1 accuracy when testing on known samples, with the proposed loss yielding the best result. The EVM has the highest accuracy when tested with unknown samples, but when we look at the accuracy for known samples, we notice that it tends to classify everything as unknown. Similar to the EVM, CROSR and CAC-OSR also classify most of the samples as unknown, leading to a high accuracy when considering just the unknowns, but very poor accuracy for known samples. Reversing this trend, OSRCI classifies most of the classes as known, which indicates it produces high prediction scores for both known and

unknown samples, but is not able to distinguish between known and unknown samples, and at the same time classifying known samples into wrong classes. SVM and OpenMax are more balanced when it comes to OSR and are reasonably accurate in detecting unknown samples, but like other algorithms, they failed to perform well when assigning known samples to the correct class. Only the MSD-Net configurations are able to maintain good performance for both novelty detection and classification.

The poor performance of the baseline approaches on the ImageNet335 OSR task was surprising. The most significant cause is likely dataset size. As emphasized earlier, OSR research tends to use a handful of small datasets with far fewer than 335 classes. Testing performance is known to drop when the number of classes increases [1]. Further, we have intentionally curated the dataset used in this paper to provide relatively few examples per class in order to assess the impact of adding behavioral data. Larger-scale OSR work has made use of Tiny ImageNet [14], [15], [21] or full ImageNet pre-training [11], [12] with access to hundreds of thousands of labeled images for training — an order of magnitude more data than what is available in our experiments for training. Our proposed approach gains information from the behavioral data that partially compensates for having fewer training samples per class, which otherwise heavily degrades performance.

Secondary reasons for poor performance include input image size sensitivity for deep learning-based models and feature size constraints for standalone classifiers. In the former case, to be able to use this data with the baselines, we had to downscale the images to match the input size required by each baseline method, which potentially led to a loss of information when they were trained. In the latter case, we had to use PCA to reduce the dimensionality of the MSD-Net features for a few of the baseline methods (details can be found in Supp. Mat. Sec. 2), available online, including: SVM, W-SVM, P_I-SVM and EVM. Such brittleness needs further attention in OSR work.

VI. Discussion and Future Work

While our method is applied to a limited number of training samples in this study, it is also possible to utilize the psychophysical loss function on larger datasets (including those with outlier samples [24]). We argue that human RT can still improve model performance, because it provides extra information for deciding whether a sample should be known or unknown. It is worth mentioning that with a larger training dataset, a larger number of RT measurements may be necessary. Limitations for the proposed method are focused around the data collection. For instance, if a large amount of human data is required then the data collection process can be very time consuming. Further, the collected human data is task specific, and is difficult to transfer to other applications. However, these are opportunities for future work.

Turning to humans as a reference point for OSR problems is a promising new direction of work, and our results hint at the potential other experimental practices from the field of psychology might have in providing information beyond simple labels to supervised learning. Gaze measurement has already been suggested as one data stream [33]. More intriguing though is the measurement of pupil dynamics. A recent study has shown that movement of the pupil can be used as an index of mental effort exertion [39], and thus could be another indicator of the patterns of error in training data.

There is also work that looks at the relationship between human behavior and machine learning models beyond RT, suggesting that human perception does not always align with model decisions. Geirhos et al. [40] suggest that humans and computer vision models process information in different ways, and Kotseruba et al. [41] argue that training CNNs with psychophysical data does not improve the performance of saliency models. This research provides a foundation for exploring what types of tasks can benefit from psychophysics and alternative ways to incorporate human behavior into algorithms. With RT, we are just scratching the surface.

Acknowledgments

The views contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the DARPA or ARO, or the U.S. Government.

References

[1] C. Geng, S.-J. Huang, and S. Chen, "Recent advances in open set recognition: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3614–3631, Oct. 2021.
[2] M. A. Sunday, A. Tomarken, S.-J. Cho, and I. Gauthier, "Novel and familiar object recognition rely on the same ability," *J. Exp. Psychol.: Gen.*, vol. 151, pp. 676–694, 2022.
[3] W. J. Scheirer, A. de Rezende Rocha, A. Sapkota, and T. E. Boult, "Toward open set recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 7, pp. 1757–1772, Jul. 2013.
[4] 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.
[5] Z.-L. Lu and B. Dosher, *Visual Psychophysics: From Laboratory to Theory*. Cambridge, MA, USA: MIT Press, 2013.
[6] W. J. Scheirer, S. E. Anthony, K. Nakayama, and D. D. Cox, "Perceptual annotation: Measuring human vision to improve computer vision," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 8, pp. 1679–1686, Aug. 2014.
[7] M. McCurrie, F. Beletti, L. Parzianello, A. Westendorp, S. Anthony, and W. J. Scheirer, "Predicting first impressions with deep learning," in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2017, pp. 999–999.
[8] S. Grieggs et al., "Measuring human perception to improve handwritten document transcription," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 6594–6601, Oct. 2022.
[9] T. E. Boult et al., "Towards a unifying framework for formal theories of novelty," in *Proc. 35th AAAI Conf. Artif. Intell.*, 2021, pp. 15047–15052.
[10] P. Kumar, A. A. Shrivastava, and S. Kong, "User guide for open world vision challenge," UMD and CMU, Tech. Rep., 2021. [Online]. Available: https://docs.google.com/document/d/19kfy77P6ahWRmDKq27hz_MxfJOTEe-d3cQkODGn8bVE/edit?usp=sharing
[11] A. Bendale and T. E. Boult, "Towards open set deep networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1563–1572.
[12] E. M. Rudd, L. P. Jain, W. J. Scheirer, and T. E. Boult, "The extreme value machine," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 3, pp. 762–768, Mar. 2018.
[13] D.-W. Zhou, H.-J. Ye, and D.-C. Zhan, "Learning placeholders for open-set recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 4399–4408.
[14] R. Yoshihashi, W. Shao, R. Kawakami, S. You, M. Iida, and T. Naemura, "Classification-reconstruction learning for open-set recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4011–4020.
[15] D. Miller, N. Sunderhauf, M. Milford, and F. Dayoub, "Class anchor clustering: A loss for distance-based open set recognition," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2021, pp. 3569–3577.
[16] G. Huang, D. Chen, T. Li, F. Wu, L. van der Maaten, and K. Q. Weinberger, "Multi-scale dense networks for resource efficient image classification," in *Proc. Int. Conf. Learn. Representations*, 2018, pp. 1–14.
[17] L. Ruff et al., "A unifying review of deep and shallow anomaly detection," *Proc. IEEE*, vol. 109, no. 5, pp. 756–795, May 2021.
[18] L. P. Jain, W. J. Scheirer, and T. E. Boult, "Multi-class open set recognition using probability of inclusion," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 393–409.
[19] W. J. Scheirer, L. P. Jain, and T. E. Boult, "Probability models for open set recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 11, pp. 2317–2324, Nov. 2014.
[20] Z. Ge, S. Demyanov, and R. Garnavi, "Generative openmax for multi-class open set classification," in *Proc. Brit. Mach. Vis. Conf.*, London, U.K., 2017, pp. 1–12.

[21] L. Neal, M. Olson, X. Fern, W.-K. Wong, and F. Li, "Open set learning with counterfactual images," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 620–635.
[22] P. Oza and V. M. Patel, "C2AE: Class conditioned auto-encoder for open-set recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2302–2311.
[23] P. Oza and V. M. Patel, "Deep CNN-based multi-task learning for open-set recognition," 2019. [Online]. Available: https://arxiv.org/abs/1903.03161
[24] D. Hendrycks, M. Mazeika, and T. Dietterich, "Deep anomaly detection with outlier exposure," in *Proc. Int. Conf. Learn. Representations*, 2019, pp. 1–18.
[25] A. R. Dhamija, M. Günther, and T. E. Boult, "Reducing network agnostophobia," in *Proc. 32nd Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 9175–9186.
[26] S. Kong and D. Ramanan, "OpenGAN: Open-set recognition via open data generation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 793–802.
[27] H. E. Gerhard, F. A. Wichmann, and M. Bethge, "How sensitive is the human visual system to the local statistics of natural images?," *PLoS Comput. Biol.*, vol. 9, no. 1, 2013, Art. no. e1002873.
[28] S. Eberhardt, J. Cader, and T. Serre, "How deep is the feature analysis underlying rapid visual categorization?," in *Proc. 30th Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 1108–1116.
[29] J. DiCarlo, D. Zoccolan, and N. Rust, "How does the brain solve visual object recognition?," *Neuron*, vol. 73, no. 3, pp. 415–434, 2012.
[30] N. V. K. Medathati, H. Neumann, G. S. Masson, and P. Kornprobst, "Bio-inspired computer vision: Towards a synergistic approach of artificial and biological vision," *Comput. Vis. Image Understanding*, vol. 150, pp. 1–30, 2016.
[31] J. K. Tsotsos, *A Computational Perspective on Visual Attention*. Cambridge, MA, USA: MIT Press, 2011.
[32] J. Duncan and G. W. Humphreys, "Visual search and stimulus similarity," *Psychol. Rev.*, vol. 96, no. 3, pp. 433–458, 1989.
[33] R. Zhang et al., "AGIL: Learning attention from human for visuomotor tasks," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 692–707.
[34] B. RichardWebster, S. E. Anthony, and W. J. Scheirer, "PsyPhy: A psychophysics driven evaluation framework for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 9, pp. 2280–2286, Sep. 2019.
[35] B. RichardWebster, S. Y. Kwon, C. Clarizio, S. E. Anthony, and W. J. Scheirer, "Visual psychophysics for making face recognition algorithms more explainable," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 263–281.
[36] O. Russakovsky et al., "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.
[37] T. G. Dietterich and A. Guyer, "The familiarity hypothesis: Explaining the behavior of deep open set methods," *Pattern Recognit.*, vol. 132, 2022, Art. no. 108931.
[38] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.
[39] P. van der Wel and H. van Steenbergen, "Pupil dilation as an index of effort in cognitive control tasks: A review," *Psychon. Bull. Rev.*, vol. 25, no. 6, pp. 2005–2015, 2018.
[40] R. Geirhos, C. R. M. Temme, J. Rauber, H. H. Schütt, M. Bethge, and F. A. Wichmann, "Generalisation in humans and deep neural networks," in *Proc. 32nd Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 7549–7561.
[41] I. Kotseruba, C. Wloka, A. Rasouli, and J. K. Tsotsos, "Do saliency models detect odd-one-out targets? New datasets and evaluations," in *Proc. Brit. Mach. Vis. Conf.*, 2019, pp. 1–15.