

Using Human Perception to Regularize Transfer Learning

Justin Dulay and Walter J. Scheirer
 Dept. of Computer Science and Engineering
 University of Notre Dame
 {jdulay, wscheire}@nd.edu

Abstract

Recent trends in the machine learning community show that models with fidelity toward human perceptual measurements perform strongly on vision tasks. Likewise, human behavioral measurements have been used to regularize model performance. But can we transfer latent knowledge gained from this across different learning objectives? In this work, we introduce **PERCEP-TL** (Perceptual Transfer Learning), a methodology for improving transfer learning with the regularization power of psychophysical labels in models. We demonstrate which models are affected the most by perceptual transfer learning and find that models with high behavioral fidelity — including vision transformers — improve the most from this regularization by as much as 1.9% Top@1 accuracy points. These findings suggest that biologically inspired learning agents can benefit from human behavioral measurements as regularizers and psychophysical learned representations can be transferred to independent evaluation tasks.

1. Introduction

All visual systems process data into latent representations before making decisions. Biological systems receive input through the eyes and encode visual inputs through the optic nerve where it can be processed within the brain through synapses and neural activity. Similarly, artificial vision systems encode image inputs into a functional mathematical space. Images are complex, labels are fuzzy, and features are noisy for many reasons. Both biological and artificial systems capture comparable internalized representations, but a high-fidelity shared representation remains elusive.

One way to capture biological latent representations of data is through psychophysics: the study of systematically changing a stimulus and measuring human response to it. In the past, studies have modified learning pipelines by incorporating some human behavioral measurements into the loss function [21], or within the model architecture itself [26]. Often, human response times towards stimuli yield

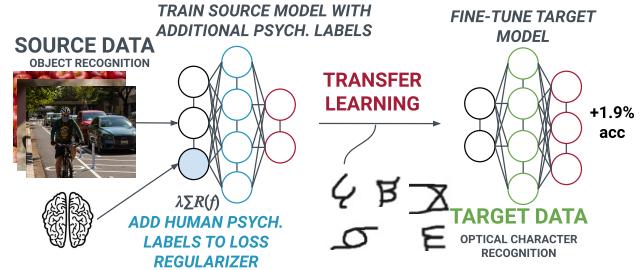


Figure 1. **PERCEP-TL** is a model agnostic transfer learning method that uses human behavioral measurements as regularization. Transfer learning, which improves with traditional regularization, often improves furthermore with psychophysical regularization. Overall, the method incorporates human behavioral measurements, such as response times, as additional labels into the regularization term of a loss function during source model training. After this, the model remains more robust to changes in target task distributions (in this case, optical character recognition, but other tasks also transfer). The target model reflects this with increased accuracy on target data and tasks.

salient information for a learning model, acting as a form of regularization. Furthermore, psychophysical measurements have been deployed for model architecture search and model evaluation in studying how a model reacts to psychophysical experiments in a similar way to a human [3, 52, 71]. This is also akin to alignment, where models that have human-like fidelity are rewarded more. A rich field of psychophysical literature exists [12, 17, 34, 39, 42, 43], including some work bridging psychophysical experiments and machine learning, but there remains a gap in using psychophysical labels beyond the scope of learning on a singular domain and task.

To efficiently share learned representations across domains and tasks, transfer learning incorporates the weights of one learned model towards a different learning objective. In recent years, applications of it have seen sweeping improvements in generalist computer vision, as well as in specialties such as medicine and robotics [6, 27, 72]. There have been recent works using regularization within transfer learning [61]. Moreover, regularization improves most transfer learning

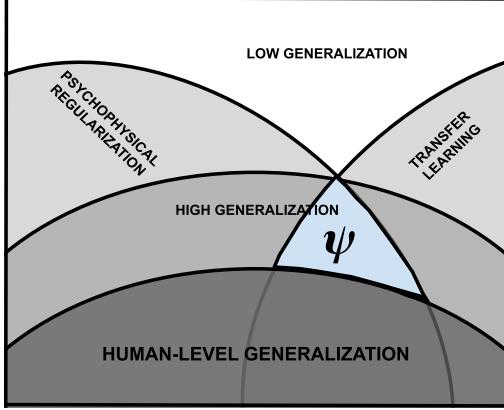


Figure 2. *What does it mean to achieve generalization?* Human-level generalization is more profound than any current artificial systems on perceptual tasks. However, many artificial systems can achieve high levels of generalization through various regularizing methods. Artificial systems can achieve better generalization when using human behavioral measurements as additional labels. Likewise, transfer learning generalizes across tasks by transferring latent knowledge from one domain to another. **Our work** combines these two artificial approaches to achieve better generalization.

tasks.

However, there is no work bridging these two core concepts: humans generalize high-fidelity representations of data that can regularize machine learning models, and transfer learning systems improve with regularization. Can we use psychophysically-annotated data to instill strong representations into models such that they can be transferred to other tasks, even *after* fine-tuning? More succinctly, can a model fine-tuned with psychophysically-labeled regularizers improve transfer learning tasks even more than a model trained with only traditional regularizers?

In this paper, we posit that psychophysical labels can help models generalize across tasks: *that when a dataset contains reliable human behavioral measurements, regularizing the loss function towards a more human-centric belief space yields improved model generalization across tasks.* Likewise, we also hypothesize that models on a spectrum of being more biologically inspired tend to also benefit proportionally more from psychophysical regularization than models that are qualitatively less biologically inspired. DiCarlo et. al [8] suggest that humans solve latent representations shared among different samples in visual recognition. A model that more closely matches the latent feature representation space of the human brain is indeed ideal, as suggested by Figure 8.

In our work, we empirically test these ideas. First, we find that models that utilize psychophysical labels within regularizers than their counterparts that do not have these special annotations. Secondly, we show that transfer learning also improves with models that use psychophysical regularization.

Likewise, we also find that models that are directly ac-

tivation fidelity, as per qualitative definitions in previous works [32, 33], often improve performatively when used in conjunction with psychophysical regularizers (though their absolute performance remains is still not as good as traditional deep learning methods). Models with **behavioral fidelity**, such as vision transformers [9] with human-like biases [16], also perform better with psychophysical regularizers.

Lastly, we compare the models on Brain-Score [54], demonstrating that several models improving *via* psychophysical labels also show stronger resemblances to state-of-the-art neural activity measurements, albeit not all models. We find that behavioral fidelity does not always correlate with activation fidelity.

Our main contributions are:

- (a) Psychophysical labels, when added to the regularization terms in loss functions, improve model generalization.
- (b) Transfer learning among different vision tasks improves when using models pre-trained with psychophysical labels and regularization.
- (c) We show Brain-Score [54] evaluations of regularized models with neural activity.

Our results show promise in transfer learning and regularization, which are fundamental concepts in machine learning research. Psychophysical studies unlock potential for more resilient learning representations in future machine learning research.¹

2. Related Work

Transfer Learning. Transfer learning is the knowledge acquisition of a source domain and task and applying it through a learned model on a target domain and task [62]. Over the past several decades, transfer learning has gained immense traction, in particular, after deep learning became fashionable [73]. With supervised learning, niche domains include medicine (*e.g.* transferring domain knowledge of radiological samples from many adult patients to few pediatric patients [5]) and simulation-to-reality transfer in robotic applications [70]. In the mainstream, many researchers use transfer learning when using an ImageNet pre-trained PyTorch model [44] or when efficiently fine-tuning BERT on different corpora [25] (among many other examples).

Similar to psychophysics, transfer learning began from experimental psychology. In 1901, Thorndike and Woodworth introduced the *transfer of practice*, that transfer from one domain to another was as good as the similarities between the domains [41]. Likewise, acclaimed psychologist B.F. Skinner considers transfer learning a form of generalization [56]. We take a similar approach in our work by complementing generalization in transfer learning with regularization from psychophysics.

¹Our code will be released after publication.

Psychophysics. Psychophysical measurements of human behavior represent a richer source of information for supervised machine learning. Many psychophysical experiments collect human perceptual responses through carefully designed experiments in response to varying stimuli. While psychophysics originated long ago [12], research in machine learning communities have demonstrated promising results incorporating ideas inspired by it into machine learning paradigms.

Scheirer et al. [53] introduced a method for incorporating reaction times of psychophysical measurements into a support vector machine loss function. In their labeling task, their crowd-sourced annotator selected items *via* an alternative forced choice task, where the response time was recorded for each action and subsequently added as an additional label for each class. These techniques have been applied in a variety of domains, such as affective recognition [36, 45], robotics [37, 69], reinforcement learning [22], and human document transcription [21], among others [4, 8, 13, 50]. Likewise, there is a growing focus in the literature to bridge gaps of robotic learning systems towards generalized intelligent agents [11, 20, 38, 48] — all point towards generalization as a concept. In our work, we use psychophysical labels as a regularizing tool; that is, if psychophysical regularizers help learning systems generalize, then they also should help transfer learning tasks.

Scaling Neural Architectures Upon Biological Fidelity. Since their inception, neural models have had various degrees of reference to biological fidelity. Convolutional neural networks are loosely inspired by biological networks, yet score high on heuristics associated with **biological fidelity** [54]. However, vision transformers, while not explicitly stated to be inspired by the human brain, have more similar biases to humans than convolutional neural networks do [16]. Likewise, other networks, such as predictive coding networks, are directly biologically-inspired with high **activation fidelity** but maintain poor performance on learning tasks. In our work, we explore how these models, on varying degrees of behavioral fidelity, perform when using them with psychophysical regularizers.

Vision transformers [9] have seen sweeping performance enhancements on benchmarks in recent years, along with some interesting comparisons to human performance on human tasks [63]. They have shown promising results on a variety of tasks including image captioning [30, 46] and feature representation [31], tasks that have been utilized in previous psychophysics and machine learning studies, among others [23].

Vision transformers appear to have different implicit biases than convolutional neural networks. Tuli et al. [63] suggested that they may be biased more towards shapes, as opposed to textures [16], which is a human trait [63] and implores higher behavioral fidelity than other methods.

Other research has also suggested that Vision Transformers' learned representations are different than convolutional neural networks and remain open to further study [40, 47, 58].

Predictive coding is a theory of learned representations in humans [49], where the brain updates a working model of its perceptual representation over time. Directly inspired by neurological predictive coding, PredNet was introduced as an artificial imitation through a series of convLSTM recurrences over time steps, encoding temporal information [32]. This model is high in activation fidelity with a direct attempt to model a psychological theory on predictive coding of neurological signals [49]. While the original task was frame prediction, object recognition [66], pose estimation [1], and object-matching have been researched [3] with it — all of which are suitable for transfer learning between tasks.

Our work marries the combination of transfer learning and psychophysics as complementary generalization components.

3. Core Concept: PERCEP-TL

In this section, we present **PERCEP-TL**, including:

- (a) How transfer learning occurs among tasks in this paper 3.1.
- (b) Psychophysical regularization for fine-tuning models 3.2.
- (c) The datasets that we used in our experiments 3.3.
- (d) Justification of psychophysical regularization in the context of transfer learning 3.4.

3.1. Transfer Learning

PERCEP-TL is the utilization of psychophysical labels in the regularization term of the loss function during model fine-tuning and transfer learning with this regularized model. This process is flexible in that it remains agnostic to initial training sets and models, meaning that, for example, a model can be trained on an object recognition task, fine-tuned with psychophysical labels, then transfer learn on a handwriting task. It trains with traditional supervised learning models can also be evaluated on unsupervised learning tasks.

More formally, **PERCEP-TL** consists of a training stage component and a transfer learning component:

- (i) Pre-train model θ_p on prior source domain \mathcal{D}_p on task \mathcal{T}_s (*optional*).
- (ii) Train a model θ_s on source domain \mathcal{D}_s on \mathcal{T}_s .
- (iii) Fine-tune θ_s with additional psychophysical labels \mathcal{Y}^ψ .
- (iv) Transfer learned model $\theta_s^\psi \rightarrow \theta_t^\psi$ on target domain \mathcal{D}_t on task on \mathcal{T}_t .

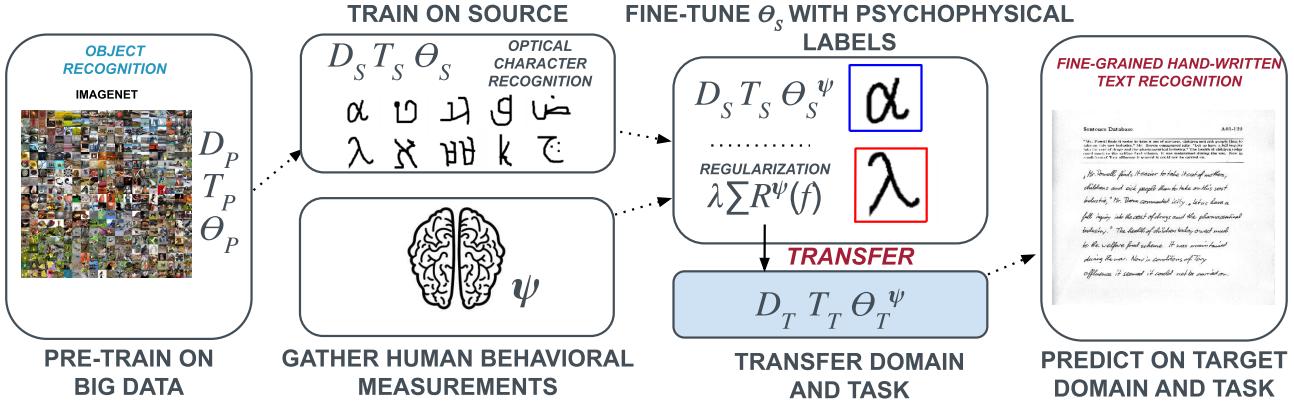


Figure 3. We visualize a neural model θ , a domain \mathcal{D} , and a learning task \mathcal{T} . In the *optional* left panel, the three components operate in the pre-training stage p . In the second column of panels, θ_s is trained on the source domain \mathcal{D}_s with source task \mathcal{T}_s , while **human behavior measurements** ψ are also gathered in the same source domain \mathcal{D}_s through **regularizing** the loss function with regularization term R . Next, θ_s^ψ is fine-tuned in the source domain with additional ψ . This latent knowledge is transferred, as in typical transfer learning, to target domain, task, and model t .

The superscript ψ represents psychophysical labels (*e.g.* reaction times) gathered from human behavioral measurements which are used in fine-tuning of the source model θ_s . With these general steps, we create a framework for transferring domain knowledge from one task to another with most models. Note, not all models need be pre-trained, but many are originally pre-trained on ImageNet. Every domain \mathcal{D} represents a dataset, and every task \mathcal{T} represents a learning task (*e.g.* multi-class classification). These steps are also outlined in Figure 3.

For example, in an object recognition task, psychophysical labels ψ are gathered from human response times to correctly identify an object from some noise. ψ is included in the regularization term of the loss function during the fine-tuning of a model θ_s on the same source object recognition data. From there, the psychophysically learned model can be used on a target task like traditional transfer learning, but with additional psychophysical regularization as **PERCEP-TL**.

3.2. Regularization with Psychophysical Labels

Preliminaries. Regularization prevents a model from being biased towards out-of-distribution samples. In explicit regularization cases, the loss function of the model is influenced by a regularization term. In general, this serves as a decorator for any simple loss function, and thus remains model agnostic.

A generic loss regularization appears as:

$$\mathcal{L} = \frac{1}{n} \left(\sum_{i=1}^n L(y_i, f(x_i)) + \lambda \sum_{i=1}^n R(y_i, x_i) \right) \quad (1)$$

In the above Equation 1, the loss output space \mathcal{L} results from the base prediction function, given some input x_i and

a ground truth label y_i at some batch step i in feature space $x \in \mathcal{X}$ and label space $y \in \mathcal{Y}$. For simplicity, L is a generic loss function. Likewise, R is a generic regularization function, remaining model agnostic as a function of the inputs and outputs. Regardless of the definition of L or R , λ is a regularization constant that constrains the loss space \mathcal{L} .

Common Regularizers. The most common forms of model regularization are ℓ_1 and ℓ_2 -regularizers. Both are used extensively within the machine learning community for their ability to create a smoother output space for models and increased generalization capacity [23, 60].

Likewise, it remains contested which to use at a given time [65]. Because of this contest, we believe that we can improve model regularization with a new regularizer that also combines human annotations.

ℓ_1 -regularization takes the mean absolute error of the *direct* outputs of the loss function f and applies the absolute difference in between terms: $\ell_1 = \sum_i^n |x_i|$. Likewise, the ℓ_2 -regularization remains similar to the aforementioned, but squares the logits: $\ell_2 = \sqrt{\sum_i^n (x_i)^2}$.

Dropout [57] removes weighted outputs for some layer in $\mathcal{L} = P(y, f(x))$. If the weights fall within a Bernoulli distribution such that they may be pruned, their answer is 1 in $k \sim \text{Bernoulli}(p)$, where p is a scalar hyperparameter for skewing the distribution.

ℓ_1 -regularization, ℓ_2 -regularization, and Dropout reduce model complexity and overfitting.

Formulation. Here, we discuss the formalization of **PERCEP-TL** in training regimes. Specifically, we use multi-class cross entropy loss with a modified generalized regularization term:

$$\mathcal{L} = - \left(\sum_j y_j \log(\hat{y}_j) + \lambda \sum_j R(w_j) \cdot \psi_j \right) \quad (2)$$

The left term is cross-entropy loss with \hat{y} as the predicted logits at j . R is a general regularization function on the logits $w_j \in \mathcal{L}$. ψ_j is a psychophysical regularization constant at data point j . Also, ψ_j is psychophysical measurement of human behavior. For example, ψ_j of a reaction time is the difference between the maximum reaction time of a human at a sample by the actual recorded reaction time at the sample. Following this, R is multiplied by ψ in cases in which the data receives a psychophysical penalty:

$$\psi_j = \begin{cases} w_j \cdot c & \text{if } \hat{y}_j \neq y_j \\ w_j & \text{otherwise} \end{cases} \quad (3)$$

with c is a constant scaled penalty within the data. R is specific in application; for example, it can be ℓ_1 .

3.3. Datasets

We used three psychophysical datasets for this paper. The first dataset uses a variant of ImageNet [7]; the second uses a variant of Omniglot [29]; the third is a variant on the IAM handwriting dataset [35]. More detailed descriptions can be found in the supplementary material.

Psych-ImageNet is a human-annotated dataset from a modification of ImageNet [7] with 293 total classes. Each data point has a psychophysical label (reaction time), class label, and ImageNet-sized (224x224) image associated with it.

Psych-Omniglot consists of a subset of the Omniglot dataset [29] with psychophysical annotations. The dataset contains images of handwritten characters from hundreds of typesets, many of which a typical crowd-sourced study participant would be unfamiliar with. The data is augmented with counterpart samples for each image with a deep convolutional generative adversarial network (DCGAN) [19] to increase intraclass variance and the sample size per class — all of which are forms of implicit regularization.

Psych-IAM is a psychophysically-annotated version of the original IAM dataset [35]. This dataset is similar to the annotations collected by Grieggs et. al in [21], but with a smaller subset of the data. In our work, we used 2,000 text lines with reaction time annotations.

Each annotation was collected for a character recognition task and word recognition task. A timer recorded the reaction time (response time) for the annotator to perform each of these tasks respectively. We report results for these three categories in the Experiments Section 4.

3.4. Why Use PERCEP-TL ?

Model Agnosticism. **PERCEP-TL** interfaces with a variety of machine learning models. In our work, we modify cross entropy loss variants and sophisticated predictive coding loss formulations, but the regularizer is compatible with other loss functions and models. It can substitute ℓ_1 -regularization in many situations. In our work, we experiment with the efficacy of this on deep convolutional neural networks, vision transformers, and deep predictive coding networks, but the regularizer can interface with any loss function that can also with a regularizer, which is in most supervised learning situations [2, 14, 23].

Generalizability. Regularizers, in general, reduce model overfitting towards spurious samples. In the case of **PERCEP-TL**, the outlier samples are ones where the model guesses incorrectly in a case where some human annotator guessed correctly with short reaction time. The penalty produced by the **PERCEP-TL** logits *increases* when the model gets *easy* examples wrong and *decreases* when the model guess correctly on *hard* examples.

Furthermore, while transfer learning improves with regularization [61], it improves more with psychophysical regularization. The results in this paper demonstrate increased generalization capacity with it.

Availability. Psychophysical annotations scale with models and datasets. In our experiments, we pre-train PredNet on KITTI [15], and evaluate this using psychophysical annotations from the **Psych-ImageNet** dataset. In the other experiments, we incorporate the annotations directly into the model training regime. This shows flexibility to (1) train the model directly and (2) evaluate different models *and* datasets. Data collection for crowd-sourced human psychophysical studies is relatively easy, as seen in subsection 3.3.

4. Experiments

In this section, we detail the experiments for **PERCEP-TL**. In 4.1, we detail the configurations of each of the models trained with psychophysical regularizers and the transfer learning tasks among them. In 4.2, we detail our ablations of different regularizers. In 4.3, we analyze these experimental results among transfer learning tasks. Lastly 4.4, we evaluate each transfer learned model with Brain-Score [54], a statistical framework that tests the biological fidelity of a model to neural activity.

4.1. Train Model Configurations

Below are the training model configurations. For each model, we selected a standard cross-entropy loss as a control measurement. Against this, we ran experiments with ℓ_1 -regularization, ℓ_2 -regularization, Dropout [57], ψ , and ψ +Dropout as experimental trials. Each trial ran the same 5 seeds, and the error bars were calculated via standard error.

CNN. In this, we configured runs with ResNet-50 [24] and VGG-16 [55], using the behavioral models as the Tuli et. al [63]. Both models utilized the PyTorch repository [44] and pre-trained on ImageNet-21k [7].

We use traditional CNN models because of their high usage in benchmark scores over the past decade in research [18], alongside previous experiments combining psychophysics and ML concepts. We further some of these approaches by showing generalizations to other models, as well.

ViT. Vision transformers (ViT) [9] are known to have biases towards shape, more so than textures like CNN’s do [63]. Because they have different inductive biases than CNNs, ViTs may be affected differently by using psychophysical labels in their loss functions that regulate their latent feature representations. In our work, we are interested in seeing whether a ViT is affected more proportionately than normal CNNs when using psychophysical labels.

In our experiments, we use the pre-trained ViT-L from HuggingFace [67]. The model is pre-trained on ImageNet-21k. We freeze the weights on the last layer and replace them with a linear classifier with the total number of classes (293) output gates. For each of these, we then apply the control cross entropy loss in conjunction with one of the experimental regularizers.

Psych-IAM Experiments. In our experiments with **Psych-IAM 2**, we utilized compared trials between a convolutional recurrent neural network *via PERCEP-TL*, and a visual transformer decoder [9, 64] encoder with the same loss modifications.

The **Psych-IAM** contains metrics on character recognition and overall word recognition. Because this dataset required different metrics than overall Top@1 accuracies, we reported resultant terms in Character Error Rate (CER) and Word Error Rate (WER), respectively, as described in 3.3.

PredNet. We use a deep predictive coding network (PredNet [32]) pre-trained on video frames from the KITTI dataset [15], on the cities and roads subset. We used the weights saved from this network for a prediction task on the human-annotated datasets.

PredNet outputs are frame-by-frame activations on the inputs. The activations from the first layer of PredNet were used to make a class prediction on the dataset. Therefore, the PredNet that we used was first pre-trained on KITTI, then evaluated as a transfer learning model on its first layer outputs to compare with the other models used in this experiment.

4.2. Regularization Ablations

Here, we present the ablations of regularizers on different neural models.

Observations. In our results, we observe that training error decreases more readily when using psychophysical labels in conjunction with regularizers vs the standard regularizers in most cases.

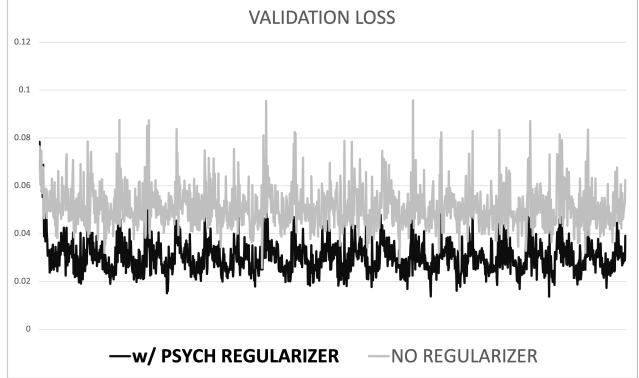


Figure 4. **Validation loss** of a PredNet model on **Psych-ImageNet**. The top curve does not use the psychophysical labels, while the bottom curve incorporates them into the regularization term R of the loss function.

Table 1 demonstrates the main results of combining psychophysical labels with standard regularizers on ResNet-50, VGG-16, ViT, and PredNet on **Psych-ImageNet** and **Psych-Omniglot** datasets. We see general trends that models fine-tuned psychophysical labels perform better than others on the same task. Below the midline, we see a row ψ and a row ψ +Dropout. The CNN models all improve more with either one of these (in some cases, the difference was minute, and both rows were bolded). Likewise, the high behavioral-fidelity model ViT also improved strongly, but the high activation-fidelity model PredNet did not show significant improvement in all but one case.

Likewise, we observe improved CER and WER on **Psych-IAM** with psychophysical regularizers when compared to control groups in Table 2. Interestingly, only the neural model improves significantly when adding a psychophysical regularization term, perhaps because the annotator labels were biased more towards a neural model.

We use these findings in transfer learning evaluation 4.3 to see how fine-tuning a model with psychophysical labels enables it to generalize better across different datasets.

Models with High Behavioral-Fidelity Improve More with Psychophysical Regularization. Often, we see the trend that the more behavioral-fidelity models have stronger proportionate changes from the additional regularization. This suggests that the salient latent information encoded by the psychophysically-annotated datasets into neural models holds better generalization capacity than datasets that do not contain these labels.

The results suggest that models with better behavioral fidelity, with improved inductive biases more *aligned* with humans, improve more from loss functions that incorporate human behavioral measurements into the regularization term.

In contrast, we see less of a gain from a psychophysical regularizer when using a high activation-fidelity model like

	Psych-ImageNet					Psych-Omniglot			
	Top@1	Top@1	Top@1	Top@1		Top@1	Top@1	Top@1	Top@1
Method	ResNet-50	VGG-16	ViT	PredNet		ResNet-50	VGG-16	ViT	PredNet
Control	0.70 ± 0.05	0.72 ± 0.05	0.74 ± 0.03	0.59 ± 0.03		0.78 ± 0.04	0.77 ± 0.05	0.81 ± 0.04	0.63 ± 0.02
ℓ_1	0.71 ± 0.01	0.72 ± 0.03	0.76 ± 0.03	0.62 ± 0.02		0.78 ± 0.03	0.78 ± 0.05	0.81 ± 0.03	0.65 ± 0.05
ℓ_2	0.72 ± 0.02	0.73 ± 0.02	0.76 ± 0.02	0.61 ± 0.03		0.77 ± 0.04	0.77 ± 0.04	0.81 ± 0.04	0.64 ± 0.03
Dropout	0.72 ± 0.02	0.73 ± 0.02	0.76 ± 0.02	0.61 ± 0.05		0.77 ± 0.04	0.77 ± 0.03	0.81 ± 0.04	0.66 ± 0.05
ℓ_1 +Dropout	0.73 ± 0.02	0.74 ± 0.05	0.79 ± 0.02	0.62 ± 0.04		0.77 ± 0.04	0.78 ± 0.02	0.83 ± 0.04	0.66 ± 0.04
ψ	0.74 ± 0.04	0.76 ± 0.03	0.78 ± 0.05	0.64 ± 0.04		0.79 ± 0.05	-	0.83 ± 0.04	0.66 ± 0.05
ψ +Dropout	0.75 ± 0.03	0.68 ± 0.08	0.80 ± 0.04	0.65 ± 0.02		0.81 ± 0.03	-	0.83 ± 0.03	0.66 ± 0.04

Table 1. In source models using psychophysical labels ψ in an additional **regularization** term, we see improved test accuracy on classification tasks averaged across seeds on the **Psych-ImageNet** and Psych-Omniglot datasets (**Psych-IAM** is pictured in Table 2). Each row denotes the regularization method utilized with multi-class cross-entropy loss. We computed error bars using standard error across 5 seeds. Columns with multiple bolded results indicate overlap between standard errors. In the psychophysical data point for VGG-16, the model always overfit the data here, yielding negligible results. VGG-16 has 138M parameters, compared to the 25M parameters for ResNet-50.

Method	Psych-IAM	
	CER	WER
ResNet-50+CRNN	0.15 ± 0.01	0.33 ± 0.02
ViT	0.13 ± 0.02	0.44 ± 0.01
ResNet-50+CRNN+ ψ	0.11 ± 0.01	0.31 ± 0.02
ViT+ ψ	0.14 ± 0.01	0.42 ± 0.02

Table 2. We perform experimental runs on the **Psych-IAM** dataset using a ResNet-50+CRNN and ViT architectures in the first two rows. In the bottom two rows, we regularize the loss function with psychophysical labels, represented by ψ . This improves both CER and WER on this dataset, suggesting further generalizability of psychophysical regularizers in fine-grained tasks. *Lower is better.*

PredNet. We further postulate that the model attunes more toward activation fidelity than toward outcome fidelity. Psychophysical labels do not possess activations — they are a distribution of human behavioral responses to similar stimuli of what the model sees, but still higher level than neuronal activity.

While these results are preliminary in that only a few models were tested in the zoo of models that exist, it shows that there may exist potential for using psychophysical regularizers to improve performance in generalizability in a variety of domains.

4.3. Transfer Learning Evaluation

In these experiments, we compared how a fine-tuned model on a psychophysically annotated dataset performed on a different dataset. We fine-tuned a ResNet-50, VGG-

16, ViT, and PredNet ² on each psychophysically annotated dataset. From there, each fine-tuned model was run on a different dataset from what it was fine-tuned with. For each trial, each model was run 5 different times on each of the transfer learning tasks, with error bars taking the standard error among each of the runs.

Figure 5 demonstrates the potential of transfer learning of a psychophysically fine-tuned model towards another dataset. Overall, models that are trained on a more variant dataset provide greater transfer learning performance gains.

In particular, models that were fine-tuned on **Psych-ImageNet**, a high-variance dataset, transferred performance well. This is likely due to the fact that **Psych-ImageNet** consists of hierarchical collections of object images; human participants, when providing annotations on this data, naturally recognize complex representations of objects more readily than most artificial systems [8, 16] (see Supp. Mat. for more details on ablations).

In contrast, most models that were fine-tuned on datasets with lower complexity than the transfer objective did not see the same improvements. For example, the ResNet-50s fine-tuned on the **Psych-IAM** handwriting tasks performed slightly worse on the **Psych-ImageNet** dataset than models which did not receive this transfer tasks (grouping 3 in Figure 5).

PERCEP-TL only remain helpful when human annotators, who handle cognitively valuable tasks, transfer to tasks of similar or less complex. Likewise, the distributions, in pixels and otherwise, between **Psych-ImageNet** and **Psych-IAM** are different. Furthermore, when flipping the transfer learning tasks, with models pre-trained on **Psych-ImageNet** against the **Psych-IAM** dataset, see a slight % increase instead of a loss.

²PredNet was pre-trained on KITTI cities first. For more details on this, please see supp. mat.

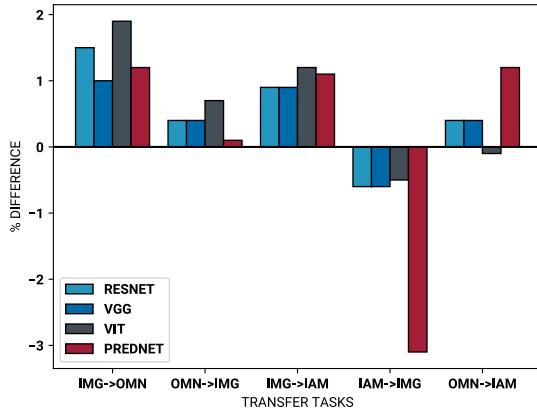


Figure 5. The % difference of average model performance across several transfer learning tasks when adding **PERCEP-TL**. Notice model performance generally improves across transfer learning tasks, but the models on the [Psych-IAM](#)→[Psych-ImageNet](#) fair much worse. The information gained from images of documents does not contain enough discriminability to transfer to a more complex dataset. Just as Skinner postulated, transfer learning is still limited by tasks with strong perceptual differences [56]. In this case, it remains impossible to generalize from fine-grained handwriting data to generic object recognition.

Qualitatively, a caveat to our method was finding the right tasks to transfer models among. We decided to keep the results transferring [Psych-IAM](#)→[Psych-ImageNet](#) to offer a counter to the claim that **PERCEP-TL** works for every transfer learning task. In addition, each transfer task would improve with further hyperparameter optimization, but we held those constant for these experimental trials for consistency.

4.4. Brain-Score Evaluation

Brain-Score is an effective metric for computing correlations among activations in biologically-inspired neural models by comparing them to neural activations in biological units, but it does not account for *behavioral* differences. Higher brain-scores correlate with biological fidelity with respect to neural activations in the visual cortex [54].

There is also recent work [51] in measuring behavioral fidelity in neural models, but it fails to have a comprehensive suite of benchmarks like Brain-Score. In further analysis, we hope to use additional metrics beyond Brain-Score once they reach consistency. Additionally, we care about explicit activation fidelity with Brain-Score, so we do not utilize perceptual similarity metrics [28, 68]. The biological mechanism may differ from the neural activity aligned with the learning task, resulting in a lower Brain-Score than expected.

This contradicts the observation that some models in our experiments saw greater gains from **PERCEP-TL** despite

Model	Brain-score
ResNet-50	0.432
ViT	0.374
PredNet	0.182
ψ +ResNet-50	0.445
ψ +ViT	0.377
ψ +PredNet	0.182

Table 3. **Brain-Score** for best models trained on [Psych-ImageNet](#). Brain-Score serves as a metric for neurological fidelity by measuring the activations of models and comparing them to neural activations in biological systems. *Higher is better.*

having a lower Brain-Score. In particular, PredNet has good transfer learning improvements when using additional psychophysical annotations, yet has an extremely low Brain-Score. We postulate that running PredNet on only one activation node, as opposed to a greater number of them, limited the amount of biologically-plausible activations for Brain-Score to measure with its task suite. These results may indicate that activation fidelity does not always correlate with behavioral fidelity.

5. Conclusion, Limitations, Broader Impacts

We introduced **PERCEP-TL** as a method to generalize transfer learning tasks. We compared the performance of a variety of machine learning models with psychophysical regularizers among ℓ_1 -regularization, ℓ_2 -regularization, and Dropout as regularizers and a control group for each model with no regularizer present. Likewise, we tested model behaviors among different transfer learning tasks. Lastly, we used Brain-Score to compare neurological fidelity among learned models. **PERCEP-TL** performs better than traditional transfer learning, and we hope this work inspires further research into psychophysical research in the computer vision community.

Any use of **PERCEP-TL** is limited by the size and fidelity of its human-annotated dataset. In our experiments, we used data collected over simple Amazon Mechanical Turk trials. While we demonstrated that the annotations gathered in this space had transferred benefits to other models, the extent of this benefit could potentially change in different types of machine learning tasks. Despite limitations, crowd-sourced data is still simple to collect when compared to tightly controlled laboratory measurements while still improving model performance. In future works, we hope to see **PERCEP-TL** in other paradigms, such as reinforcement learning.

Because **PERCEP-TL** remains easily scales without revealing much about the annotator, we do not anticipate adverse effects with it alone, but we caution for adversarial use cases that further prejudice and unjust biases.

References

- [1] Jamal Banzi, Isack Bulugu, and Zhongfu Ye. Learning a deep predictive coding network for a semi-supervised 3d-hand pose estimation. *IEEE/CAA Journal of Automatica Sinica*, 7(5):1371–1379, 2020. 3
- [2] Raef Bassily, Om Thakkar, and Abhradeep Guha Thakurta. Model-agnostic private learning. *Advances in Neural Information Processing Systems*, 31, 2018. 5
- [3] Nathaniel Blanchard, Jeffery Kinnison, Brandon Richard Webster, Pouya Bashivan, and Walter J Scheirer. A neurobiological evaluation metric for neural network model search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5404–5413, 2019. 1, 3, 13
- [4] Aidan Boyd, Patrick Tinsley, Kevin Bowyer, and Adam Czajka. Cyborg: Blending human saliency into the loss improves deep learning. In *Proceedings of the IEEE/CVF Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, 2023. 3
- [5] Vikash Chouhan, Sanjay Kumar Singh, Aditya Khamparia, Deepak Gupta, Prayag Tiwari, Catarina Moreira, Robertas Damaševičius, and Victor Hugo C De Albuquerque. A novel transfer learning based approach for pneumonia detection in chest x-ray images. *Applied Sciences*, 10(2):559, 2020. 2
- [6] S Deepak and PM Ameer. Brain tumor classification using deep cnn features via transfer learning. *Computers in Biology and Medicine*, 111:103345, 2019. 1
- [7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 5, 6
- [8] James J DiCarlo, Davide Zoccolan, and Nicole C Rust. How does the brain solve visual object recognition? *Neuron*, 73(3):415–434, 2012. 2, 3, 7
- [9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 2, 3, 6
- [10] Justin Dulay, Sonia Poltoratski, Till S Hartmann, Samuel E Anthony, and Walter J Scheirer. Guiding machine perception with psychophysics. *arXiv preprint arXiv:2207.02241*, 2022. 12, 13
- [11] Alireza Fathi, Yin Li, and James M Rehg. Learning to recognize daily actions using gaze. In *Proceedings of the IEEE/CVF Conference on European Conference on Computer Vision*, pages 314–327. Springer, 2012. 3
- [12] G. T. Fechner. Elements of psychophysics, 1860. In W. Dennis, editor, *Readings in the History of Psychology*, pages 206–213. Appleton-Century-Crofts, 1948. 1, 3
- [13] Thomas Fel, Ivan Felipe, Drew Linsley, and Thomas Serre. Harmonizing the object recognition strategies of deep neural networks with humans. *Advances in Neural Information Processing Systems*, 2022. 3
- [14] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pages 1126–1135, 2017. 5
- [15] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 5, 6, 13
- [16] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019. 2, 3, 7
- [17] George A Gescheider. *Psychophysics: the fundamentals*. Psychology Press, 2013. 1
- [18] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016. 6
- [19] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27, 2014. 5, 12
- [20] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 3
- [21] Samuel Grieggs, Bingyu Shen, Greta Rauch, Pei Li, Jiaqi Ma, David Chiang, Brian Price, and Walter Scheirer. Measuring human perception to improve handwritten document transcription. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 1, 3, 5, 12
- [22] Suna Sihang Guo, Ruohan Zhang, Bo Liu, Yifeng Zhu, Dana Ballard, Mary Hayhoe, and Peter Stone. Machine versus human attention in deep reinforcement learning tasks. *Advances in Neural Information Processing Systems*, 34:25370–25385, 2021. 3
- [23] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. A survey on vision transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 3, 4, 5
- [24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 6
- [25] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR, 2019. 2
- [26] Jin Huang, Derek Prijateln, Justin Dulay, and Walter Scheirer. Measuring human perception to improve open set recognition. *arXiv preprint arXiv:2209.03519*, 2022. 1, 12, 13
- [27] Benjamin Q Huynh, Hui Li, and Maryellen L Giger. Digital mammographic tumor classification using transfer learning from deep convolutional neural networks. *Journal of Medical Imaging*, 3(3):034501, 2016. 1
- [28] Manoj Kumar, Neil Houlsby, Nal Kalchbrenner, and Ekin Dogus Cubuk. Do better imagenet classifiers assess perceptual

- similarity better? *Transactions of Machine Learning Research*, 2022. 8
- [29] Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015. 5, 12
- [30] Wei Liu, Sihan Chen, Longteng Guo, Xinxin Zhu, and Jing Liu. Cptr: Full transformer network for image captioning. *arXiv preprint arXiv:2101.10804*, 2021. 3
- [31] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 3
- [32] William Lotter, Gabriel Kreiman, and David Cox. Deep predictive coding networks for video prediction and unsupervised learning. In *International Conference on Learning Representations*, 2017. 2, 3, 6, 13
- [33] William Lotter, Gabriel Kreiman, and David Cox. A neural network trained to predict future video frames mimics critical properties of biological neuronal responses and perception. *Nature Machine Learning*, 2018. 2
- [34] Vera Maljkovic and Ken Nakayama. Priming of pop-out: I. role of features. *Memory & Cognition*, 22(6):657–672, 1994. 1
- [35] U-V Marti and Horst Bunke. The iam-database: an english sentence database for offline handwriting recognition. *International Journal on Document Analysis and Recognition*, 5(1):39–46, 2002. 5, 12
- [36] Mel McCurrie, Fernando Beletti, Lucas Parzianello, Allen Westendorp, Samuel Anthony, and Walter J Scheirer. Predicting first impressions with deep learning. In *Proceedings of the IEEE/CVF Conference on Face and Gestures*, pages 518–525, 2017. 3
- [37] Michael Milford, Sam Anthony, and Walter Scheirer. Self-driving vehicles: Key technical challenges and progress off the road. *IEEE Potentials*, 39(1):37–45, 2019. 3
- [38] Kyle Min and Jason J Corso. Integrating human gaze into attention for egocentric activity recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1069–1078, 2021. 3
- [39] Ken Nakayama and Gerald H Silverman. Serial and parallel processing of visual feature conjunctions. *Nature*, 320(6059):264–265, 1986. 1
- [40] Muhammad Muzammal Naseer, Kanchana Ranasinghe, Salman H Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. *Advances in Neural Information Processing Systems*, 34:23296–23308, 2021. 3
- [41] Pedro Tamesis Orata. The theory of identical elements, being a critique of thorndike’s theory of identical elements and a re-interpretation of the problem of transfer of training. 1928. 2
- [42] Alice J O’Toole, Dana A Roark, and Hervé Abdi. Recognizing moving faces: A psychological and neural synthesis. *Trends in Cognitive Sciences*, 6(6):261–266, 2002. 1
- [43] Alice J O’Toole, Kenneth A Deffenbacher, Dominique Valentin, Karen McKee, David Huff, and Hervé Abdi. The perception of face gender: The role of stimulus structure in recognition and classification. *Memory & cognition*, 26(1):146–160, 1998. 1
- [44] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, pages 8024–8035. Curran Associates, Inc., 2019. 2, 6
- [45] Víctor Ponce-López, Baiyu Chen, Marc Oliu, Ciprian Corneanu, Albert Clapés, Isabelle Guyon, Xavier Baró, Hugo Jair Escalante, and Sergio Escalera. Chalearn LAP 2016: First round challenge on first impressions-dataset and results. In *ECCV Workshops*, pages 400–418, 2016. 3
- [46] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 3
- [47] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in Neural Information Processing Systems*, 34:12116–12128, 2021. 3
- [48] Santhosh K Ramakrishnan, Tushar Nagarajan, Ziad Al-Halah, and Kristen Grauman. Environment predictive coding for embodied agents. *arXiv preprint arXiv:2102.02337*, 2021. 3
- [49] Rajesh PN Rao and Dana H Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1):79–87, 1999. 3
- [50] Brandon RichardWebster, Samuel E Anthony, and Walter J Scheirer. Psyphy: A psychophysics driven evaluation framework for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(9):2280–2286, 2018. 3
- [51] Brandon RichardWebster, Anthony DiFalco, Elisabetta Caldesi, and Walter J Scheirer. Perceptual-score: A psychophysical measure for assessing the biological plausibility of visual recognition models. *arXiv preprint arXiv:2210.08632*, 2022. 8
- [52] Brandon RichardWebster, So Yon Kwon, Christopher Clarizio, Samuel E Anthony, and Walter J Scheirer. Visual psychophysics for making face recognition algorithms more explainable. In *Proceedings of the IEEE/CVF European Conference on Computer Vision*, pages 252–270, 2018. 1
- [53] Walter J Scheirer, Samuel E Anthony, Ken Nakayama, and David D Cox. Perceptual annotation: Measuring human vision to improve computer vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(8):1679–1686, 2014. 3, 12

- [54] Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J Majaj, Rishi Rajalingham, Elias B Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, et al. Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, page 407007, 2020. 2, 3, 5, 8
- [55] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015. 6
- [56] Burrhus Frederic Skinner. *Contingencies of reinforcement: A theoretical analysis*, volume 3. BF Skinner Foundation, 2014. 2, 8
- [57] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014. 4, 5
- [58] Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your vit? data, augmentation, and regularization in vision transformers. *arXiv preprint arXiv:2106.10270*, 2021. 3
- [59] Neil Stewart, Jesse Chandler, and Gabriele Paolacci. Crowd-sourcing samples in cognitive science. *Trends in Cognitive Sciences*, 21(10):736–748, 2017. 12
- [60] Shiliang Sun, Zehui Cao, Han Zhu, and Jing Zhao. A survey of optimization methods from a machine learning perspective. *IEEE Transactions on Cybernetics*, 50(8):3668–3681, 2019. 4
- [61] Masaaki Takada and Hironori Fujisawa. Transfer learning via ℓ_1 regularization. *Advances in Neural Information Processing Systems*, 33:14266–14277, 2020. 1, 5
- [62] Lisa Torrey and Jude Shavlik. Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pages 242–264. IGI global, 2010. 2
- [63] Shikhar Tuli, Ishita Dasgupta, Erin Grant, and Thomas L Griffiths. Are convolutional neural networks or transformers more like human vision? *Cognitive Science*, 2021. 3, 6
- [64] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017. 6
- [65] Colin Wei, Sham Kakade, and Tengyu Ma. The implicit and explicit regularization effects of dropout. In *International Conference on Machine Learning*, pages 10181–10192. PMLR, 2020. 4
- [66] Haiguang Wen, Kuan Han, Junxing Shi, Yizhen Zhang, Eugenio Culurciello, and Zhongming Liu. Deep predictive coding network for object recognition. In *International Conference on Machine Learning*, pages 5266–5275. PMLR, 2018. 3
- [67] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierrick Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, Oct. 2020. Association for Computational Linguistics. 6
- [68] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 8
- [69] Ruohan Zhang, Zhuode Liu, Luxin Zhang, Jake A Whritner, Karl S Muller, Mary M Hayhoe, and Dana H Ballard. AGIL: Learning attention from human for visuomotor tasks. In *Proceedings of the IEEE/CVF European Conference on Computer Vision*, pages 663–679, 2018. 3
- [70] Wenshuai Zhao, Jorge Peña Queralta, and Tomi Westerlund. Sim-to-real transfer in deep reinforcement learning for robotics: a survey. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 737–744. IEEE, 2020. 2
- [71] Sharon Zhou, Mitchell Gordon, Ranjay Krishna, Austin Narcomey, Li F Fei-Fei, and Michael Bernstein. Hype: A benchmark for human eye perceptual evaluation of generative models. *Advances in Neural Information Processing Systems*, 32, 2019. 1
- [72] Zhuangdi Zhu, Kaixiang Lin, and Jiayu Zhou. Transfer learning in deep reinforcement learning: A survey. *arXiv preprint arXiv:2009.07888*, 2020. 1
- [73] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020. 2

Abstract

Here, we discuss additional details to the main **PERCEP-TL** paper. We organize it by:

- *Dataset descriptions A.*
- *PredNet implementation details B.*
- *Ablations of other experiments C.*

The items in this supplementary material serve to provide additional detail and context to the main text while not adding or removing pertinent information from it.

A. Dataset Descriptions.

A.1. Psych-ImageNet

- The dataset has 293 known classes in total, excluding other open-set classes to use at later studies.
- There are 40 classes with psychophysical labels, producing a ratio of psychophysically-annotated to original classes as in [53].
- There are 33,548 known training samples in total, and 12,428 samples have corresponding reaction times.
- Each data point has a reaction time, class label, and ImageNet-sized (224x224) image associated with it.
- Reaction times collected (each reaction time is the amount of time to choose a stimulus, given 5 other examples). Responses in this data were collected for class recognition against noisy stimuli.

Each trial was an object-matching task, where the human participant of 5 opposed stimuli to it (see supplementary material). Each image was from one of the 293 Psych-ImageNet classes chosen for the task. The participant had to select the object they thought belonged to the top sample or rejected it together, should there be no match. A timer collected the participants reaction time for each question. Best viewed in color. An example of crowd-sourced task pairings from [26] is shown in Figure 6.

Classes were evenly distributed across trials, as were positive vs negative matches. Likewise, the difficulty of experiments was variable to avoid a ceiling effect, a form of scale attenuation in which the maximum performance measured does not reflect the true maximum of the independent variable. More dataset details can be found in the supplementary material.

A.2. Psych-Omniglot

The Psych-Omniglot is a variant on the Omniglot dataset [29] with psychophysical labels collected from the

research in [10]. The dataset contains images of handwritten characters from hundreds of typesets, many of which a typical crowd-sourced study participant would be unfamiliar with. The data is augmented with counterpart samples for each image with a deep convolutional generative adversarial network (DCGAN) [19] to increase intraclass variance and the sample size per class — all of which are forms of implicit regularization. An example of crowd-sourced task pairings from [10] is shown in Figure 7.

In this dataset, human behavioral measurements were gathered as reaction times to stimuli in crowd-sourced experiments. Human participants were presented with two opposing stimuli from the original Omniglot dataset (a Two-Alternative Forced Choice task) and decided whether the two stimuli were the same character in the dataset. The reaction time from the participants was recorded automatically. Broadly speaking about the dataset as a whole, these human reaction times were long on hard pairings, and short on easy character pairings. The introduction of this easy vs. hard pairing would prove useful for supervised learning tasks.

A.3. Psych-IAM

The dataset is a modification of the IAM dataset [35] with human behavioral measurements on lines of text collected from [21] on about 35% of the dataset (2,152 lines).

In the main text, we report both word error rate (WER) and character error rate (CER) for this dataset. The word error rate is a model’s error with respect to the individual word on the line in the dataset, while the character error rate corresponds to the model’s fidelity with the human annotator’s marks on the individual word.

The reaction time of the annotator to accurately record a character and line was recorded in this annotated dataset [21]. For the main text study, we used the reaction times in conjunction with images of the text itself to perform transfer learning OCR tasks with **PERCEP-TL**.

A.4. Dataset Limitations

We recognize that **Psych-ImageNet** only contains annotations on 40 of the total 293 classes. While previous psychophysics and machine learning studies suggest that this still remains representative of the entire training distribution [21, 53], reaction times on more classes may potentially yield better results.

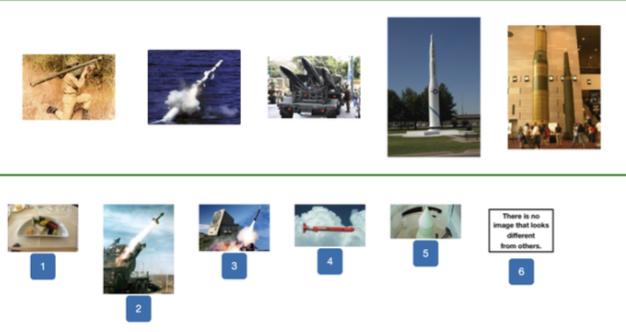
Likewise, Psych-Omniglot is a dataset in which the annotations were collected *via* Amazon Mechanical Turk. While the practitioners accounted for systematic errors, no crowdsourcing study is entirely robust to untrustworthy annotators [59].

Lastly, **Psych-IAM**, along with the other two datasets, also suffer from limits of overall numbers of available annotations due to academic budget constraints.

In spite of limitations, we show ?? that even with a

Experiment: 1/25

In the bottom row, when looking at the images from left to right, which image is the first that is different from those in the upper row?



Experiment: 0/25

Which of these images in the second row is most different from the others?



Figure 6. Crowd-sourced tasks from [26]. The above figure contains two screenshots from the worker data aggregator view in Amazon Mechanical Turk. The image on the left contains an **easy** example where most annotators answered quickly and accurately; a model that fails to answer in the same way receives a higher penalty. Likewise, the screenshot on the right contains a more **difficult** class, where a model does not receive as harsh of a penalty for answering incorrectly.



Figure 7. Crowd-sourced tasks from [10]. An example two-alternative forced choice OCR task as seen from the participant’s view. Labels (d) and (f) represent character pairs where the class labels differ; the rest represent the same class pairing. The blurred and noisy images lead to more informative psychophysical labels for operationalization within the machine learning task during training.

small level of annotations, model performance still generally improves.

B. PredNet Fine-Tuning.

Formulation. psychophysical transfer learning pulls similar results on the evaluative steps, as well.

Similarly to Blanchard et al. [3], we extract the activations of PredNet after the convLSTM layers. PredNet works with temporal data. First, We pre-trained it on videos from the KITTI [15] self-driving dataset. We set up the annotated Psych-ImageNet in an order of fixed frames and record the activations of PredNet at the fixed time steps. This representation at each neuron works like a supervised model’s neuron in that we can add psychophysical transfer learning

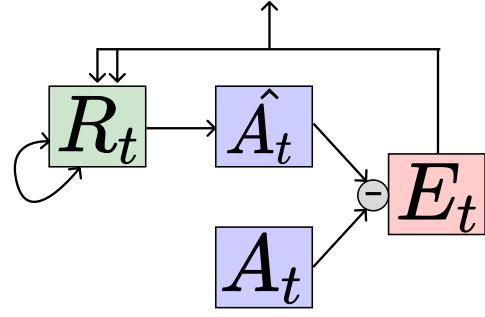


Figure 8. The PredNet [32] architecture.

to it to regularize the learning representation.

The loss defined by PredNet is as follows:

$$\mathcal{L}_{train} = \sum_t \lambda_t \sum_l \frac{\lambda_l}{n_l} \sum_{n_l} E_l^t \quad (4)$$

where λ_t is a regularizer at the time step, $\frac{\lambda_l}{n_l}$ is a regularizing factor at a given layer in the network, and E_l^t is the error at a time step 8.

Indeed, the loss formulation for PredNet is inherently more complex than cross-entropy loss variations. For brevity, we conducted experiments to understand in which term should the psychophysical regularization variables be used. Again, each of the three terms uses a form of ℓ_1 -normalization to adjust model learning generalization.

We observe that multiplying psychophysical transfer learning into the layer term λ_t — with the variable \hat{A}_t yields the best results. In other words, after successive outputs of each layered convLSTM, we see performance gains more vividly than any other term within this loss. Furthermore, the regularization effect of psychophysical transfer learning,

Psych-ImageNet	MAE
Method	PredNet
Control	0.59 ± 0.03
ℓ_1	0.62 ± 0.02
ℓ_2	0.61 ± 0.03
Dropout	0.61 ± 0.05
Dropout+ ℓ_1	0.61 ± 0.02
<i>RegularPsych</i>	0.64 ± 0.04
<i>RegularPsych</i> +Dropout	0.65 ± 0.02

Table 4. On models using *RegularPsych* as an evaluator, we see improved mean squared error reduction. All models were pre-trained on KITTI and evaluated on the house dataset. We computed error bars using standard error across 5 seeds. *Lower is better.*

Psych-ImageNet	Psych-Omniglot	
Parameter	Train Error	Train Error
None	0.12 ± 0.03	0.20 ± 0.05
A_t	0.11 ± 0.04	0.19 ± 0.05
\hat{A}_t	0.06 ± 0.02	0.17 ± 0.04

Table 5. The table shows the train errors for each parameter selection of which PredNet architecture layer to multiply by the *RegularPsych* variable. The None column assumes a cross-entropy loss without any modification to the PredNet loss. The input layer A_t shows no significant change in performance, regardless of what the psychophysical annotations are. However, we see a significant reduction in training error when applying *RegularPsych* to the model prediction logits \hat{A}_t .

the softening of sharp gradient turns, pronounces the most at longer time steps on average (*e.g.* at $steps > 5$). As table 4 suggests, the loss mostly benefited from psychophysical transfer learning regularization on the outputs between step outputs \hat{A}_t at times t 5.

This result suggests that predictive coding networks in some way manage the latent ideals encoded in the psychophysical transfer learning data.

While this experiment step was not part of the *model-evaluative*, we believed it important to fine-tune psychophysical transfer learning on a non-traditional loss framework before conducting experimentation on the relative effects of psychophysical transfer learning on model-evaluative performance.

The pre-training of PredNet and the subsequent transfer to task to a frame-by-frame prediction on the modified Psych-ImageNet allows for the beneficial usage of psychophysical transfer learning. While this case is a niche, it demonstrates the viability of utilizing psychophysical transfer learning in a variety of future neurologically-inspired models.

C. Model Ablations

In Table 6, we report ablation results on the **PERCEP-TL** transfer learning tasks. These show some additional transfer learning movements among different tasks in the experiments. For example, the first row of the figure represents the task shift, where the color of the ψ represents the domain the psychophysical labels were gathered on. Not all domains transfer well, but there exist several domains where transfer learning works naturally.

In this work, it remains apparent that the object recognition task and psychophysical labels from models learned on **Psych-ImageNet** transfer well to the other domains used in this study. In rows 1 and 3 in Table 6, we see the largest gains supported by this. Likewise, the transfer of domains from **Psych-IAM** character annotation tasks to generic object recognition, in line with commonsense, does not transfer well.

In future studies, we plan to explore different learning paradigms (*e.g.* reinforcement learning) to expand the results of transfer among domains.

Transfer Task	orig. + new + %diff			orig. + new + %diff			orig. + new + %diff			orig. + new + %diff		
	ResNet			VGG			ViT			PredNet		
$\psi \rightarrow \psi$	0.79	0.81	+1.5%	-			0.83	0.85	+1.9%	0.63	0.65	+1.2%
$\psi \rightarrow \psi$	0.74	0.75	+0.4%	0.76	0.76	+0.4%	0.78	0.79	+0.7%	0.65	0.65	+0.1%
$\psi \rightarrow \psi$	0.91	0.92	+0.9%	0.81	0.02	-0.5%	0.86	0.88	+1.2%	0.64	0.65	+1.1%
$\psi \rightarrow \psi$	0.74	0.73	-0.6%	0.76	0.76	+0.2%	0.78	0.77	-0.5%	0.65	0.62	-3.1%
$\psi \rightarrow \psi$	0.91	0.91	+0.4%	0.81	0.81	+0.1%	0.86	0.86	-0.1%	0.65	0.66	+1.2%

Table 6. **Transfer learning % difference table.** With psychophysical transfer learning, performance increases by as much as **1.9%**. Each row represented in this table represents a difference transfer learning task, denoted by ψ corresponding in color with the dataset used. Each trial is the standard error across 5 seeds. *Note: using accuracy as $1 - CER$ on *Psych-IAM* trials in this table. Higher is better.*