Do better ImageNet classifiers assess perceptual similarity better?

Anonymous authors
Paper under double-blind review

Abstract

Perceptual distances between images, as measured in the space of pre-trained deep features, have outperformed prior low-level, pixel-based metrics on assessing image similarity. While the capabilities of older and less accurate models such as AlexNet and VGG to capture perceptual similarity are well known, modern and more accurate models are less studied. In this paper, we present a large-scale empirical study to assess how well ImageNet classifiers perform on perceptual similarity. First, we observe a inverse correlation between ImageNet accuracy and Perceptual Scores of modern networks such as ResNets, EfficientNets, and Vision Transformers: that is better classifiers achieve worse Perceptual Scores. Then, we examine the ImageNet accuracy/Perceptual Score relationship on varying the depth, width, number of training steps, weight decay, label smoothing, and dropout. Higher accuracy improves Perceptual Score up to a certain point, but we uncover a Pareto frontier between accuracies and Perceptual Score in the mid-to-high accuracy regime. We explore this relationship further using a number of plausible hypotheses such as distortion invariance, spatial frequency sensitivity, and alternative perceptual functions. Interestingly we discover shallow ResNets, trained for less than 5 epochs only on ImageNet, whose emergent Perceptual Score matches the prior best networks trained directly on supervised human perceptual judgements.

1 Introduction

ImageNet (Russakovsky et al., 2015) is the cornerstone of modern supervised learning and has enabled significant progress in computer vision. Features learnt via training on ImageNet transfer well to a number of downstream tasks (Carreira & Zisserman, 2017; Zhai et al., 2019; Huang et al., 2017), making ImageNet pretraining a standard recipe. Further, better accuracy on ImageNet usually implies better performance on a diverse set of downstream tasks such as robustness to common corruptions (Orhan, 2019; Xie et al., 2020b; Taori et al., 2020; Radford et al., 2021), adversarial robustness (Cubuk et al., 2017; Xie et al., 2020a), out-of-distribution generalization (Recht et al., 2019; Andreassen et al., 2021; Miller et al., 2021), transfer learning on smaller classification datasets (Kornblith et al., 2019), pose estimation (Mathis et al., 2021), domain adaptation (Zhang & Davison, 2020), object detection and segmentation (Zoph et al., 2020), and for predicting neural recordings and behaviors of primates on object recognition tasks (Schrimpf et al., 2020).

As a remarkable side effect, ImageNet models also encode abstract, higher-level representations from images. Distances measured in the intermediate feature space of ImageNet models capture a notion of similarity identical to humans. The first generation of ImageNet classifiers: AlexNet (Krizhevsky et al., 2012), VGG (Simonyan & Zisserman, 2014), and SqueezeNet (Iandola et al., 2016) can all measure perceptual similarity, termed as Perceptual Scores (PS), in a way that outperforms all prior pixel-level metrics and correlates better with human judgement (Zhang et al., 2018). However, while image classification has progressed significantly since then, the ability of modern ImageNet networks to assess perceptual similarity is unknown.

Motivated by this, in this paper, we attempt to answer the question: Are better ImageNet classifiers better in assessing perceptual similarity? We perform a suite of experiments on BAPPS, a large dataset of human-evaluated perceptual judgements (Zhang et al., 2018). To the best of our knowledge, our work is the first

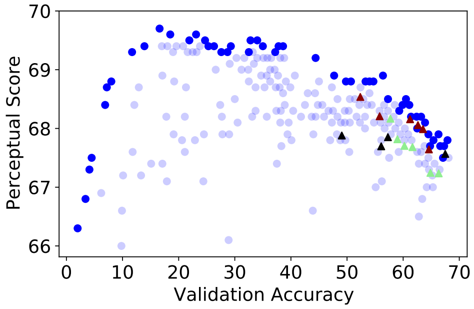


Figure 1: We discover a Pareto frontier (marked in dark blue) between Perceptual Scores (Zhang et al., 2018) on the 64 × 64 BAPPS Dataset (y-axis) and ImageNet 64 × 64 validation accuracies (x-axis). Each blue dot represents an ImageNet classifier. Better ImageNet classifiers achieve better Perceptual Scores up to a certain point. Beyond this point, improving on accuracy hurts Perceptual Score. Modern networks, EfficientNets (▲), Vision Transformers (▲) and ResNets (▲) lie far to the right of the point of optimal Perceptual Score, achieving high accuracy and lower Perceptual Scores. The best Perceptual Scores are attained by classifiers with moderate accuracy (20.0-40.0). The lowest and highest Perceptual Score observed in prior networks are 64.3 (using random weights, i.e. an untrained network) and 68.9 (AlexNet) respectively.

empirical study to present a systematic investigation and rigorous deep dive into the relationship between ImageNet accuracy and perceptual similarity. We study the ImageNet accuracy/PS interplay of a wide variety of networks across many combinations of architectures and hyperparameters.

While Zhang et al. (2018), show a positive correlation between PS and ImageNet accuracy, we observe that this holds only in the low-accuracy regime. In the mid-to-high accuracy regime, we uncover a counter-intuitive Pareto frontier between accuracy and PS. Contrary to prevailing evidence that suggests models with high validation accuracies on ImageNet are likely to transfer better on a downstream task, we find that representations from underfit ImageNet models with modest validation accuracies achieve the best PS. Our experiments further suggest that attaining the best PS is somewhat architecture agnostic. For example, ResNets trained with large amounts of weight decay or early stopped within a few epochs of training can match or outperform the PS of AlexNet and VGG.

We explore the relationship between ImageNet accuracy and PS using a number of plausible hypotheses. While the causatory latent factor that governs the relationship between accuracy and perceptual similarity remains unclear, our paper opens the door to further understanding of this phenomenon.

A summary of our experiments are:

- We systematically evaluate the PS of modern “out-of-the-box” networks, ResNets, EfficientNets and Vision Transformers.
- We study the variation of PS (and accuracy) as a function of width, depth, number of training steps, weight decay, label smoothing and dropout. Our large-scale study consists of 722 different ImageNet networks across the cross product of these 7 different hyperparameters and 5 architectures.
- We explore the relationship between ImageNet accuracy and PS further using spatial frequency sensitivity, invariance to distortions, class granularity, and improved global perceptual functions.

Our empirical study leads to the following surprising results:

- While modern classifiers outperform prior pixel-based metrics in PS, they under-perform moderate classifiers like AlexNet.

- Modern ImageNet models that are much shallower, narrower, early-stopped within a few epochs of training and trained with larger values of weight decay attain significantly higher PS than their out-of-the-box counterparts.
- In all of our hyperparameter sweeps with the exception of label smoothing and dropout, we discover an unexpected and previously unobserved tradeoff between accuracy and PS. In each hyperparameter sweep, there exists an optimal accuracy up to which improving accuracy improves PS. This optimum is fairly low and is attained quite early in the hyperparameter sweep. Beyond this point, better classifiers achieve worse PS.
- Perceptual functions that rely on global image representations such as style achieve better PS than per-pixel perceptual functions.
- We do not find a correlation between PS of models and their sensitivity to distortions. Further, low Perceptual Score models are not necessarily more reliant on high-frequency spatial information for classification as compared to high PS models.
- High-level features found in the latter layers of residual networks, have better PS than low-level features found in the earlier layers.
- Finally, we find particularly shallow, early-stopped ResNets trained only on ImageNet that attain an emergent Perceptual Score of 70.2. This matches the best Perceptual Scores across prior networks which were trained directly on BAPPS to match human judgements.

2 Related Work

There has been a rich body of work dedicated to analyzing the transfer of pretrained ImageNet networks to various downstream tasks. Kornblith et al. (2019) show that transfer learning performance is highly correlated with ImageNet top-1 accuracy. Taori et al. (2020); Miller et al. (2021) show that improvements in accuracy on ImageNet consistently results in improvements on test datasets with distribution shift. Djolonga et al. (2021) show a positive correlation between ImageNet transfer performance and out-of-distribution robustness. While Geirhos et al. (2018) demonstrate that ImageNet-trained models have a stronger bias towards using texture cues compared to humans, Hermann et al. Hermann et al. (2019) show that among high-performing models, shape-bias is correlated with ImageNet accuracy. Schrimpf et al. (2020) found that CNNs that achieve a higher accuracy on ImageNet are better at predicting neural recordings of primates when executing object recognition tasks, but the positive correlation was weaker for high-accuracy models. While better ImageNet classifiers transfer better on the above tasks, we are the first to observe a negative correlation between ImageNet classifiers and their inherent ability to capture image similarity. While we do expect that larger models with more capacity can attain better PS when trained directly on BAPPS, we use BAPPS solely to evaluate the emergent perceptual properties of ImageNet-trained classifiers.

An orthogonal line of work, involves incorporating "perceptual losses" that minimize the distance between intermediate features of a ground-truth and generated image to drive photorealistic synthesis in style transfer (Gatys et al., 2015), (Li et al., 2017), super-resolution (Johnson et al., 2016), and conditional image synthesis (Dosovitskiy & Brox, 2016), (Chen & Koltun, 2017). Following the investigation done in (Zhang et al., 2018), state-of-the-art models in image-synthesis domains such as super-resolution (Lugmayr et al., 2020; 2021), (Bhat et al., 2021), image inpainting (Nazeri et al., 2019), (Zhao et al., 2021), high-resolution image synthesis (Esser et al., 2021), (Karras et al., 2019), (Karras et al., 2020), deblurring (Kupyn et al., 2019), (Zhang et al., 2020), image-to-image translation (Richardson et al., 2021), (Lee et al., 2018b), video generation (Babaeizadeh et al., 2021), (Franceschi et al., 2020), (Wu et al., 2021), (Lee et al., 2018a), neural radiance fields (Martin-Brualla et al., 2021), (Mildenhall et al., 2020), view synthesis (Wang et al., 2021), (Riegler & Koltun, 2020), (Tucker & Snavely, 2020), (Wiles et al., 2020), and others (Lai et al., 2018), (Liu et al., 2021) use perceptual distance as either an auxiliary training loss to improve image synthesis or as an evaluation metric to assess synthesized image quality. While there has been some anecdotal evidence regarding the greater efficacy of VGG features over ResNet-50 for style losses (Nakano, 2019), there has been no systematic investigation assessing the relationship between ImageNet accuracy and perceptual similarity. In this paper, we exclusively focus on uncovering the relationship between the accuracy of ImageNet models and their inherent ability to capture perceptual similarity and not on applications of perceptual losses to downstream tasks.

3 Background: Perceptual Score

We first define the PS of a network, for which we adopt the "2AFC" scoring protocol (Zhang et al., 2018). First, let x_0 and x_1 denote two images. Let $\tilde{y_0}^l$ and $\tilde{y_1}^l$ be the feature maps for x_0 and x_1 at the lth layer of a network, normalized across the channel dimension. The perceptual similarity function $d(x_0, x_1)$ is defined as:

$$d(x_0, x_1) = \sum_{l \in \mathcal{L}} \frac{1}{H_l W_l} \sum_{h,w} ||\tilde{y_0}^l_{h,w} - \tilde{y_1}^l_{h,w}||^2 \tag{1}$$

where H_l and W_l denote the height and width of the feature maps at layer l, respectively. $\mathcal{L}$ denotes the subset of layers which are used in the perceptual similarity; this subset is architecture specific. Given a reference image x and two target images x_0 and x_1, BAPPS (Zhang et al., 2018) provides a ground truth soft-label p. p can be interpreted as the probability a human rater would rate x_1 as more similar to x than x_0.

For a neural network with distances $d_0 = d(x, x_0)$ and $d_1 = d(x, x_1)$, from (Zhang et al., 2018) we define the PS $s(d_0, d_1)$ to be the following value times 100:

$$p\mathbb{1}[d_0 > d_1] + (1-p)\mathbb{1}[d_1 > d_0] + 0.5\mathbb{1}[d_0 = d_1] \tag{2}$$

Dynamic Range. Unlike accuracy that can vary from 0 to 100, PS have a narrow dynamic range. The previous best PS for low-level metrics was attained by FSIMc (Zhang et al., 2011), a low-level hand crafted matching function with a value of 63.8. The previously reported lower and upper PS bounds for ImageNet networks (Zhang et al., 2018) were values of 64.3 and 68.9. These were achieved by a randomly initialized network and AlexNet (Krizhevsky et al., 2012) respectively. The human upper bound is 73.9. Across all our experiments, we cover almost the entire prior dynamic PS range of trained networks, with the PS of our networks ranging from 65.0 to 69.7.

4 Experimental Setup

Architectures. We study two of the most popular classes of vision architectures: Convolutional Neural Networks (CNNs) and Transformers. As representative CNNs, we train ResNets (He et al., 2016) and EfficientNets (Tan & Le, 2019). For EfficientNets, we use model variants B0 to B5 (Tan & Le, 2019). For ResNets, we train networks with depths ranging from 6 to 200 layers. For Vision Transformers (ViT) (Dosovitskiy et al., 2020), we use the Base and Large variants, with patch sizes of 8 and 4, leading to four models (ViT-B/8, ViT-B/4, ViT-L/8 and ViT-L/4). We use smaller patch sizes than is common because we use lower resolution images (see the following Training section). However, any patch size smaller than 4 leads to poor generalization.

Training. We train our networks on ImageNet at a resolution of 64 × 64 and report their accuracies on the ImageNet validation set and PS on BAPPS. ImageNet 64 × 64 provides a cleaner test bed instead of the standard high-resolution (224 × 224 and above) for the following reasons.

BAPPS is constructed using 64 × 64 images so that human raters can focus on low-level local changes as opposed to high-level semantic differences. ImageNet models pre-trained on high-resolution images classify 64 × 64 images poorly. Such models are sub-optimal for our analysis, because we are interested in the perceptual properties of classifiers that generalize reasonably well. An alternative option would be to resize

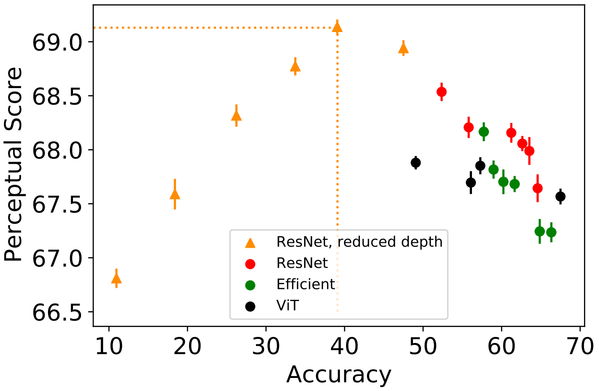


Figure 2: Perceptual Scores (Zhang et al., 2018) on the 64 × 64 BAPPS dataset as a function of ImageNet 64 × 64 validation accuracies. Each point is the average across 5 runs. Error bars on both the axes are the corresponding standard errors (variance in accuracy is sufficiently small that error bars in the x-direction are barely visible). Across the out-of-the-box ResNets, EfficientNets and ViTs (circles), there exists a negative correlation between Perceptual Score and accuracy. On reducing the depth even further, (ResNet-reduced depth, triangles) we uncover an "optimal accuracy threshold". Below this threshold, Perceptual Scores correlate with accuracy and above it Perceptual Scores decrease. Among the architectures in this plot, ResNet-6 attains this optimal accuracy threshold and has the highest Perceptual Score.

64×64 images to high-resolution images using linear interpolations. However, this can have the adverse side effect of blurring out or removing distortions from BAPPS images.

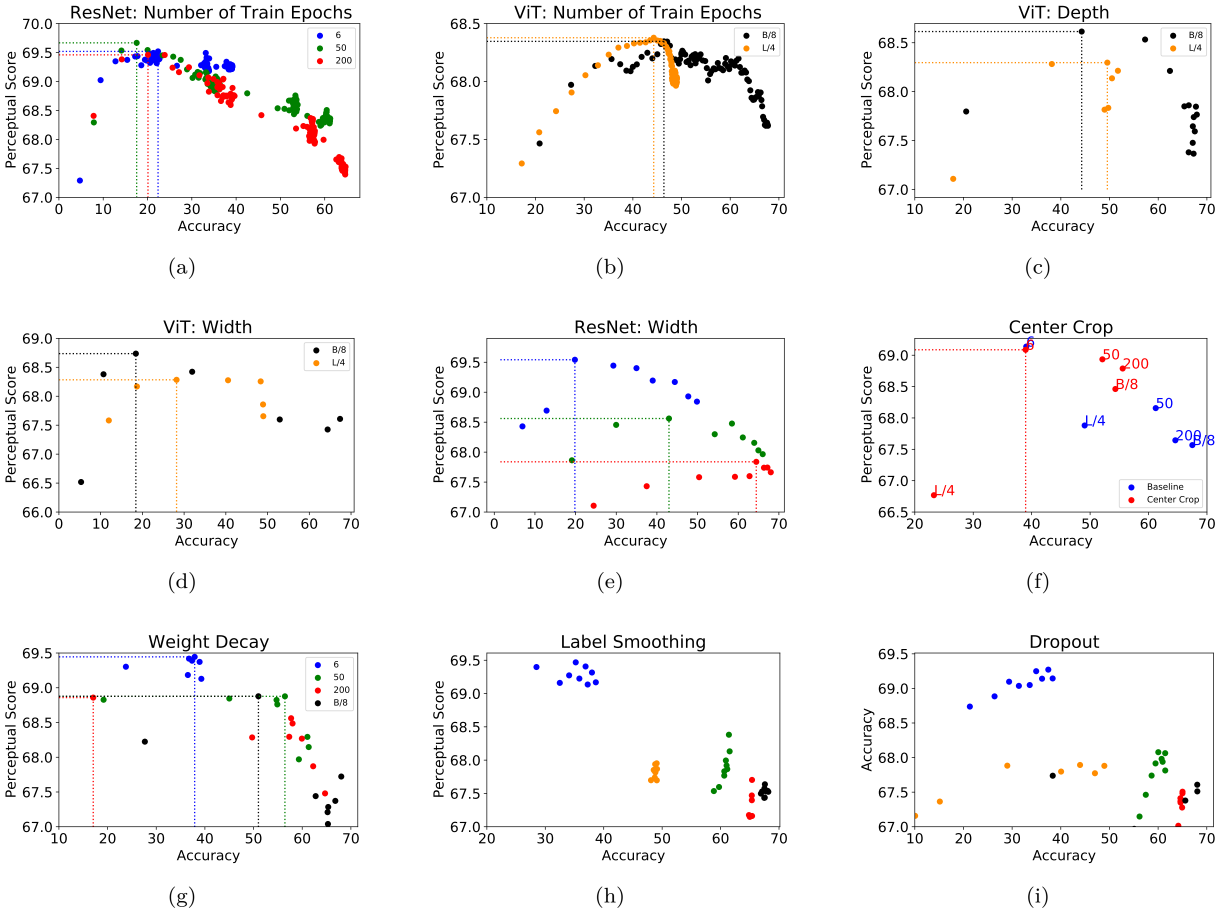
For all networks, we apply only random crops and flips and disable all other augmentations. We use the recommended hyperparameter settings given by the corresponding open-sourced training code. For the models where validation accuracy decreases during training, EffNet-B4 and EffNet-B5, we stop training just before this happens.

Representations. Here, we describe which representations are used to compute PS for each architecture. For ResNets and EfficientNets, we use the inputs to the reduction stages, where the spatial dimensions are reduced via strided convolutions. Specifically, we obtain four 2D representations of resolution 16×16, 8×8, 4×4, and 2×2. For Vision Transformers, we use the global CLS representation of the image at the output of every encoder block. Our initial experiments on using features at the input of every layer instead of every block or reduction stage lead to worse PS across all architectures.

5 ImageNet accuracy versus Perceptual Score

We train out-of-the-box networks five times and report the mean accuracy and PS (64 × 64) with error bars in Fig. 2. All modern networks (red, green, and black) in Fig. 2 obtain a higher PS than FSIMc and randomly initialized networks. Surprisingly though, representations of better ImageNet networks perform consistently worse. For example, ResNet-18 which achieves the best PS of 68.5 has a modest accuracy of 52.0% (64 × 64). EfficientNet B5 which has the worst PS of 67.2 achieves a much higher accuracy of 66.3%. Additionally, all networks perform worse than A lexNet.

Next, we push this observation to the limits. Our next experiment hopes to identify a much smaller network that achieves improved PS at the expense of extremely low accuracy. In Fig. 2, we report the accuracies and PS obtained by shallow ResNets with depths from 2 to 10 (ResNet, reduced depth). ResNet-10 and ResNet-6 obtain improved PS with reduced accuracies. However, the ResNets with a depth smaller than 6 incur losses in both accuracy and PS. Our results indicate that better classifiers produce better perceptual representations up until a certain "optimal accuracy“. Above this optimal accuracy, ImageNet classifiers trade off better accuracies with worse PS. This phenomenon leads to some interesting observations: 1)



(a) (b) (c) (d) (e) (f) (g) (h) (i)

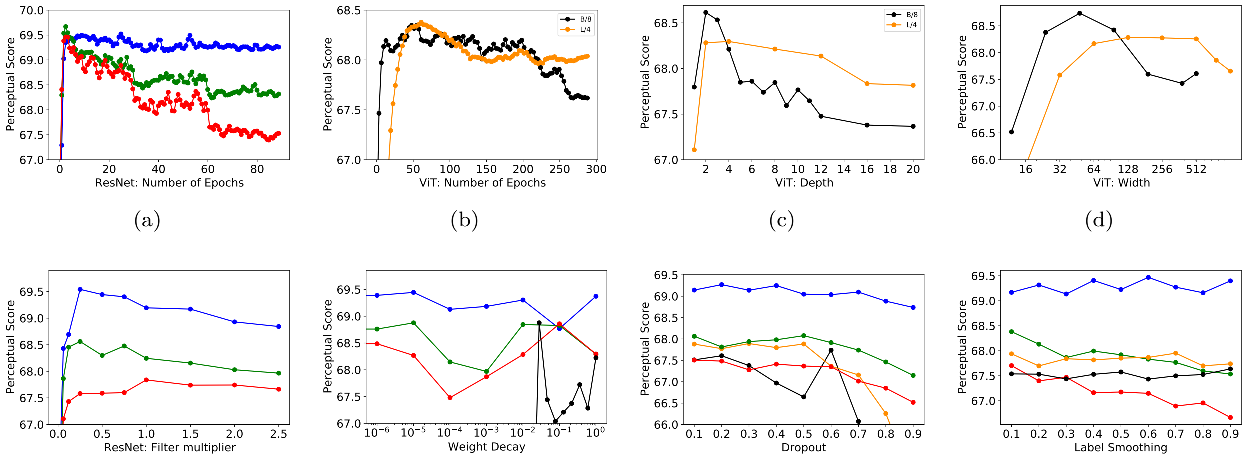
Figure 3: The relationship between Perceptual Scores and ImageNet validation accuracy when varying different hyperparameters. Each plot depicts the ImageNet accuracy (x axis) and Perceptual Score (y axis) obtained by a 1D sweep over that particular hyperparameter.

ResNet-6 happens to achieve this optimal accuracy with a PS of 69.1, outperforming AlexNet. 2) Beyond the optimal accuracy, there is a strong inverse correlation between accuracy and PS (coefficient = -0.84). 3) ResNet-200 and ResNet-3 achieve similar PS, while having a 45% difference in accuracy.

6 How general is this relationship?

Section 5 indicates that an inverse-U relationship exists between accuracy and PS on varying the depth of residual networks. In this section, we attempt to generalize this relationshop as a function of other implicit hyperparameters such as layer composition, width, depth and training hyperparameters of the considered architectures. In particular, we assess this relationship between accuracy and PS as a function of via 1D hyperparameter sweeps. We vary a single hyperparameter of a network (e.g. width) along a 1D grid which changes the network's accuracy and as an effect, the corresponding PS. For a given sweep, we provide two sets of plots: 1) Scatter plots between PS/accuracy (Fig. 3). 2) PS against the hyperparameter values ((Fig. 4). We choose 5 architectures: ResNets and ViTs with the best and worst PS, (ResNet-6, ResNet-200, ViT-B/8, ViT-L/4) plus a standard ResNet-50.

Figs. 3 and 4 summarize our experiments across the following hyperparamters: Number of training steps, network width, network depth, weight decay (Hanson & Pratt, 1988), dropout (Srivastava et al., 2014), and label smoothing (Szegedy et al., 2016b). We define p_{max} to be the optimal PS and $a(p_{max})$ to be the accuracy



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 4: In each plot, we vary a single hyperparameter along a 1D grid and plot the Perceptual Scores.

at which p_{max} is reached. For the hyperparameter settings, where we observe the inverse-U relationship, we indicate p_{max} and $a(p_{max})$ in Fig. 3 with dotted lines parallel to the x and y axes, respectively.

Number of train epochs. All networks exhibit the inverse-U shaped behaviour (Fig. 3a and 3b). $a(p_{max})$ is fairly consistent within a given family of architectures. ResNets have $a(p_{max})$ at low values between 15-25 % and ViTs have $a(p_{max})$ at moderate values between 40-50 %.

PS peaks early during training across all networks (Fig. 4b and 4a). ResNet-50 and ResNet-200 peak at the first few epochs of training ($p_{max} = 69.7$) while the ViT models peak at lower values ($p_{max} = 68.4$) and later around 60 epochs. After the peak, PS of better classifiers decrease more drastically. ResNets are trained with a learning rate schedule that causes a step-wise increase in accuracy as a function of training steps. Interestingly, in Fig. 4a, they exhibit a step-wise decrease in PS that matches this step-wise accuracy increase.

Width and Depth. In Fig. 3e, we see that ResNet-6 achieves $p_{max} = 69.6$ at $a(p_{max}) = 20\%$ while ResNet-200 has $p_{max} = 67.6$ at $a(p_{max}) = 65\%$. As the model capacity is increased from ResNet-6 to ResNet-200, the peak shifts from the top left to the bottom right. Compound scaling is a model scaling technique (Tan & Le, 2019; Szegedy et al., 2016a) that improves model accuracy efficiently by scaling the width and depth of a network together. We observe a peculiar "inverse compound scaling" phenomena; depth and width of networks have to be scaled down simultaneously to improve on PS significantly.

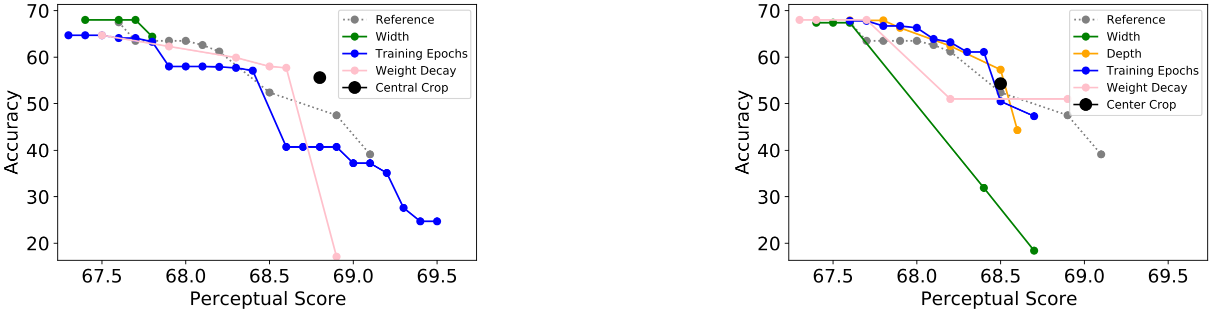
$a(p_{max})$ on varying the width and depth of ViTs are at 20% to 30% and 45-50%, respectively. (Figs. 3c, 3d). A shallow ViT model of depth 2 gets close to 40% accuracy. Hence, there might just not be enough points between 20% to 40% in Fig. 3c. This could explain the shift of $a(p_{max})$ to the right in Fig. 3c when compared to Fig. 3d.

Shallower and narrower architectures perform better as shown in Figs. 4d, 4c, and 4e. The optimal width of ViT-B/8 and ViT-L/4 are 6 and 12% of their default widths while their optimal depths are just 2 transformer blocks. ResNet-6 and ResNet-50 exhibit similar properties with the optimal width being 25% of their original widths. ResNet-200 is the outlier with a small peak at its original width.

Central Crop. Modern networks employ random crops of high-resolution rectangular images during training. This artificially increases the effective quantity of training data available to the network. Replacing random crops with central crops has been shown to increase shape bias (Hermann et al., 2019) and reduce the discrepancy of object scales between training and testing (Touvron et al., 2019).

Table 1: Perceptual Score improves by scaling down ImageNet models. Each value denotes the improvement obtained by scaling down a model across a given hyperparameter over the model with default hyperparameters.

Model	Default	Width	Depth	Weight Decay	Central Crop	Train Steps	Best
ResNet-6	69.1	+0.4	-	+0.3	0.0	**+0.5**	69.6
ResNet-50	68.2	+0.4	-	+0.7	+0.7	**+1.5**	69.7
ResNet-200	67.6	+0.2	-	+1.3	+1.2	**+1.9**	69.5
ViT B/8	67.6	+1.1	+1.0	**+1.3**	+0.9	+1.1	68.9
ViT L/4	67.9	+0.4	+0.4	-0.1	-1.1	**+0.5**	68.4



(a) ResNet-200

(b) ViT B/8

Figure 5: Each curve in a subplot showcases the validation accuracy vs Perceptual Score Pareto frontier for that architecture/hyperparameter combination. Each point on a line denotes the minimum accuracy loss achievable for a particular PS gain. The dashed gray line is the reference Pareto frontier obtained with out-of-the-box architectures.

Fig. 3f shows the accuracies and PS of the 5 architectures trained with center crops. Each architecture moves towards the top-left with lower accuracies and higher PS. ViT-L/4 is the exception as it moves towards the bottom-right. It encounters a significant reduction in accuracy, lowering it below $a(p_{max})$, which could explain the decrease in its PS. All center-cropped architectures lie along an inverse-U, with ResNet-6 at the optimum.

Weight Decay. Fig. 4f displays the impact of weight decay on PS. ViT-L/4 has minimal variation in accuracy as the weight decay factor is varied; so we omit ViT-L/4 from this study. ResNets and ViTs achieve their worst PS at the default weight decay around 10^{-4} and 0.3, respectively. The PS increases on either side of this optimum. This correlates with their changes in accuracy as a function of weight decay. ResNets and ViTs achieve their best accuracies at these weight decay values and decrease in either direction.

Label Smoothing and Dropout. Across all our controlled settings, label smoothing and dropout are the only hyperparameters that decrease both PS and accuracy.

Varying label smoothing produces less drastic changes in accuracy as compared to other hyperparameters. In Fig. 3h, a very weak positive correlation exists between accuracy and PS for ResNet-50 and ResNet-200. In Fig. 4h, PS of ResNet-50 and ResNet 200 decrease with more label smoothing, while the Vision Transformers and ResNet-6 are almost invariant. Our results indicate that clean labels are necessary to obtain high PS. However, varying label smoothing does not change accuracy a great deal within each architecture class, so the dynamic range is insufficient to observe the inverse-U relationship.

In Fig. 4g the PS consistently decreases as a function of dropout and in Fig. 3i, it also correlates with accuracies. Curiously, dropout is the only factor to negatively influence both accuracy and PS simultaneously.

Conclusion. In Fig. 1, we plot the accuracy and PS from all our above experiments. While their exact relationship is architecture and hyperparameter dependent, we uncover a global Pareto frontier between PS and accuracies, see Fig.. 1. Up to a certain peak, better classifiers achieve PS and beyond this peak, better accuracy hurts PS.

7 Scaling down improves Perceptual Scores

Our results in Section 6 prescribe a simple strategy to make an architecture's PS better: Scale down the model to reduce its accuracy till $a(p_{max})$.

Table 1 summarizes the improvements in PS obtained by scaling down each model across every hyperparameter. With the exception of ViT-L/4, across all architectures, early stopping yields the highest improvement in PS. In addition, early stopping is the most efficient strategy as there is no need for an expensive grid search.

How much does PS improvement cost in terms of accuracy? Fig. 5 shows the accuracy-PS Pareto frontier for ResNet-200 and ViT-B/8. Each point on the Pareto frontier denotes the maximum possible achievable accuracy for a given PS. The gray line is the reference Pareto-frontier obtained from training networks with their default settings. Except for Width + ViT-B/8 that lies below the reference Pareto frontier, all curves lie very close to the reference Pareto frontier. Early-stopping any of the architectures to improve their PS, as seen in Table 1, greatly reduces their accuracy.

8 Further Exploration

8.1 Global Similarity Functions

We first posit that the perceptual similarity function is suboptimal, and an alternative would not yield an inverse-U relationship.

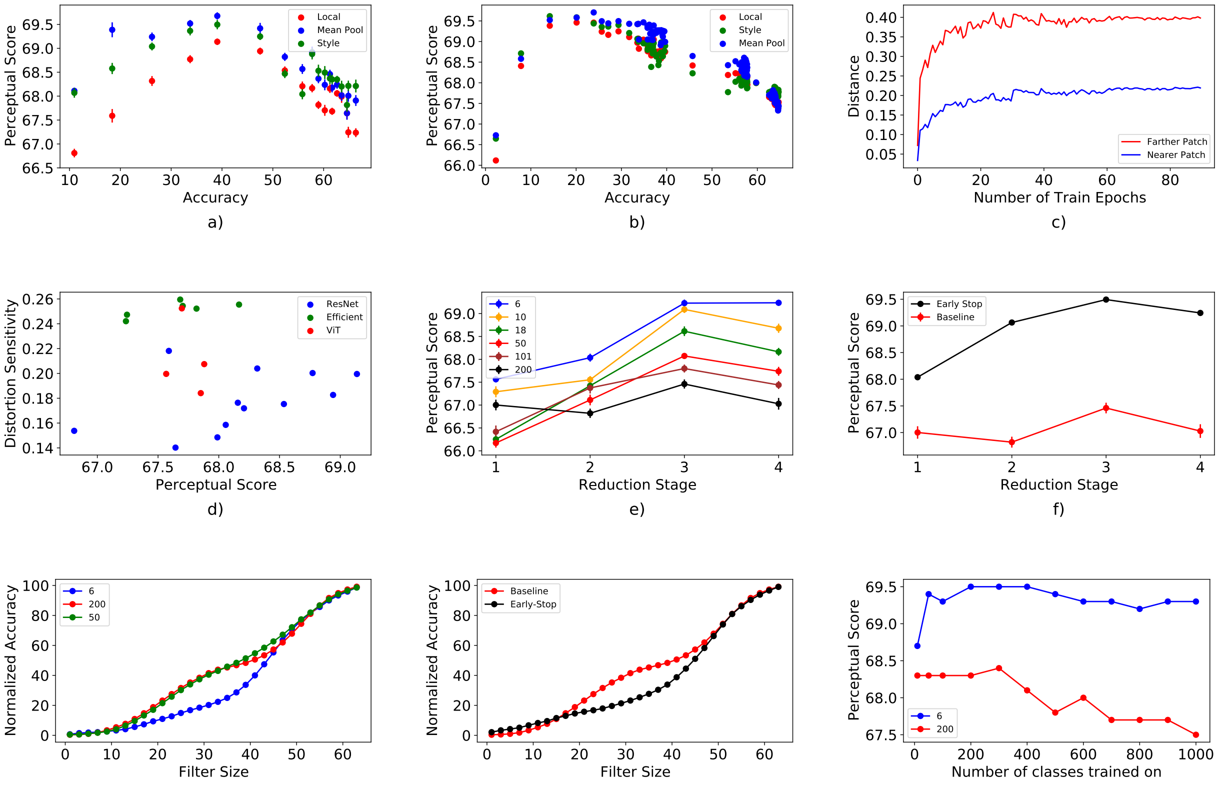
The perceptual similarity function in (Zhang et al., 2018) averages per-pixel differences across the spatial dimensions of the image. This assumes a direct correspondence between pixels, which may not hold for warped, translated or rotated images. For a similarity function that compares global representations of images, the inverse-U relationship may no longer exist. We investigate two such functions in two different settings: 1) Out-of-the-box ResNets and EfficientNets. 2) ResNet-200 as a function of train steps.

Style. We adopt the style-loss function from the Neural Style Transfer work of (Gatys et al., 2015). Let $\tilde{y}_0^l$ and $\tilde{y}_1^l \in \mathbb{R}^{H_l \times W_l \times C_l}$ be the spatially normalized outputs of two images x_0 and x_1 at the lth layer of the network. We compute $G_0^l \in \mathbb{R}^{C_l \times C_l}$, the inter-channel cross-correlation matrix of $\tilde{y}_0^l$ and similarly G_1^l for $\tilde{y}_1^l$. G_0 and G_1 also known as Gram matrices capture global information in $\tilde{y}_0^l$ and $\tilde{y}_1^l$ respectively. The style function is given by $\sum_l \frac{1}{C_l^2} ||G_0^l - G_1^l||_F^2$.

Mean Pool. Let y_0^l and $y_1^l \in \mathbb{R}^{H_l \times W_l \times C_l}$ be the unnormalized outputs at the lth layer of the network. We spatially average them to global representations and then normalize across channels to obtain $\bar{y}_0^l$ and $\bar{y}_1^l \in \mathbb{R}^{C_l}$. The distance function is then given by $\sum_l \frac{1}{C_l} ||\bar{y}_0^l - \bar{y}_1^l||^2$.

In Figs. 6a and 6b, we present scatter plots between accuracy and PS with the style and mean pool similarity functions. Both functions yield better PS than the baseline ("Local") in (Zhang et al., 2018). ResNet-6 with its Mean Pool and Style variants outperform the baseline (local) 69.1 with scores of 69.7 and 69.5 respectively. On early-stopped ResNet-200, the mean pool and style functions improve upon the baseline score of 69.5 with 69.8 and 69.7 respectively.

Early-stopped ResNet-6 with its mean pool variant achieves 70.2. This matches the best reported PS in (Zhang et al., 2018), where the AlexNet model is trained from scratch on the BAPPS train set. Note that neither of our networks have seen the BAPPS train set during ImageNet training.



a) b) c)

d) e) f)

g) h) e)

Figure 6: We attempt to demystify the relationship between accuracy and Perceptual Scores via a number of hypotheses. **Global Similarity Functions:** (Fig. 6a: Out-of-the-box Networks and Fig. 6b: ResNet-200 on varying Train Epochs) **Distortion Sensitivity:** (Fig. 6c: Out-of-the-box Networks and Fig. 6d: ResNet-200 as a function of train steps) **Layer-wise Perceptual Scores:** (Fig. 6e: ResNets and Fig. 6f ResNet-200 - Baseline (converged) vs early-stopped) **Spatial frequency sensitivity:** (Fig. 6h ResNet-200 - Baseline (converged) vs early-stopped and Fig. 6g Out-of-the-box ResNets) **Class Granularity:** (Fig. 6i Perceptual Score vs Number of ImageNet classes trained on)

However, although the improved perceptual functions attain better PS as compared to the baseline, the inverse-U correlation is still prominent. Therefore, we can conclude that while the per-pixel comparison function is suboptimal, it is not the main cause of the inverse correlation between accuracy and PS.

Learned. Lastly we investigate what happens if the similarity function is learned on supervised data. Although the main goal of our paper is to assess the inherent perceptual properties of ImageNet models, we may also train a linear layer on top of pretrained ImageNet features to match supervised human judgements on BAPPS. The PS gap between ResNet-6 and ResNet-200 narrows down from 1.5 to 0.8, but even after training, the ResNet-6 still outperforms the ResNet-200. See the appendix for more details.

8.2 Distortion Sensitivity

Here, we explore whether sensitivity to distortions is the common latent factor influencing both PS and accuracy. Intuitively, better networks will be less sensitive to the distortions introduced in the BAPPS dataset, since the class will not change under these distortions. This intuition is supported by results that show that accuracy under distribution shifts (including artificial corruptions) correlates strongly with "clean"

ImageNet accuracy (Taori et al., 2020). Therefore, if decreased sensitivity is related to poorer PS, due to inability to distinguish different class-preserving perturbations, then this could explain our observations.

Let x_f and x_n be the farther and nearer patch to a reference image x, respectively. For this experiment, we only include examples, (x, x_f, x_n), where all human raters unanimously agree that x_f is the farther patch. Concretely, in Eq 2, these are datapoints where $p = 1.0, x_f = x_0, x_n = x_1$ or $p = 0.0, x_f = x_1, x_n = x_0$.

We measure distortion sensitivity using the distance margin $\mathbb{E}_{x,x_f} d(x, x_f) - \mathbb{E}_{x,x_n} d(x, x_n)$. We expect this margin to be larger for a distortion sensitive network. In Fig. 6c, among out-of-the-box classification networks, there exists no positive correlation between distortion sensitivity and PS. As another experiment, in Fig. 6d, we plot $\mathbb{E}_{x,x_f} d(x, x_f)$ (Farther Patch) and $\mathbb{E}_{x,x_n} d(x, x_n)$ (Nearer Patch) as a function of training epochs (ResNet-200). From Fig. 4a, we know that PS decreases as a function of epochs after it reaches a peak. However in Fig. 6d, the distance margin between the farther and nearer patch remains fairly constant as a function of epochs. **Hence, low PS models are not necessarily less sensitive to distortions.**

8.3 Layer-wise Perceptual Scores

Remember, PS is averaged over many layers, see Eq 2. However, it might be the case that optimal features for PS are buried in specific layers for better classifiers (e.g. lower layers), while other layers (e.g. high layers) exhibit different behaviour more optimal for classification. Therefore, we look at the best PS across layers.

We define the "layer-wise PS" $d(x_0^l, x_1^l)$ to be $\frac{1}{H_l W_l} \sum_{h,w} ||\tilde{y_0}^l_{h,w} - \tilde{y_1}^l_{h,w}||^2$. The PS in Eq 2 can be viewed as a mean ensemble of "layer-wise PS". The "optimal layer-wise PS" is then $\max_{l \in \mathcal{L}} d(x_0^l, x_1^l)$ Fig. 6e, displays the layer-wise PS for each of the 4 2D representations across different ResNet depths. The optimal l for all depths is 3, and larger depths attain worse PS even at this optimal l. We additionally see that in Fig. 6f, ResNet-200 under-performs the optimal layer-wise PS of its early-stopped variant at $l = 3$. **Therefore, we can conclude that sub-optimal features are not a cause of the inverse-U relationship.**

8.4 Spatial Frequency Sensitivity

Networks that rely more on high-frequency information for classification could be less robust to high-frequency distortions or removal of high frequencies from an image (Yin et al., 2019), and as an effect have low PS. We analyze the relationship between spatial frequency sensitivity of different networks and their PS. A low-pass square filter of side r filters out the high frequencies in an image outside a square with edge length r in its Fourier spectrum. We measure the "normalized accuracy", which is the accuracy on low-pass filtered images divided by its accuracy on clean images as a function of r. A model more reliant on high frequency information will have a higher "normalized accuracy" slope at high values of r. ResNet-6 has a higher "normalized accuracy" slope at a high $r = 40$ to 50 as compared to ResNet-6 (Fig. 6g). Despite being more reliant on higher spatial frequencies, ResNet-6 achieves a higher PS compared to ResNet-200. Similarly, ResNet 200 becomes more reliant on higher spatial frequencies if it is early stopped (Fig. 6h), while also increasing its PS (Fig. 4a). **These results indicate that models that have low PS are not necessarily more reliant on high frequency information for classification.**

8.5 ImageNet Class Granularity

ImageNet is a 1000 class classification problem that includes fine-grained classes. A classifier that models such classes successfully could have a reduced PS, since it could compromise on learning general features. The low accuracy of ResNet-6 implies that its capacity is sufficient to model only a subset of these classes, and its inability to model tougher classes might explain its high PS.

We create subsets having number of classes ranging from 50 to 900 and train ResNet-6 and ResNet-200 networks on each of these subsets. In Fig. 6i, the PS gap between ResNet-200 and ResNet-6 reduces as the number of classes are decreased. But, ResNet-200 still underperforms ResNet-6. **Therefore, class granularity cannot fully explain why a less-accurate ResNet-6 significantly outperforms ResNet-200 on PS.**

9 Conclusion

We show that an inverse-U relationship exists between accuracy and PS across a number of settings. By scaling down networks and using more optimal perceptual functions, we match the PS of best prior networks explicitly trained to match human judgements. Finally, we also explore a number of explanations for this inverse-U relationship.

References

Anders Andreassen, Yasaman Bahri, Behnam Neyshabur, and Rebecca Roelofs. The evolution of out-of-distribution robustness throughout fine-tuning. *arXiv preprint arXiv:2106.15831*, 2021.

Maria Attarian, Brett D Roads, and Michael C Mozer. Transforming neural network visual representations to predict human judgments of similarity. *arXiv preprint arXiv:2010.06512*, 2020.

Mohammad Babaeizadeh, Mohammad Taghi Saffar, Suraj Nair, Sergey Levine, Chelsea Finn, and Dumitru Erhan. Fitvid: Overfitting in pixel-level video prediction. *arXiv preprint arXiv:2106.13195*, 2021.

Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Deep burst super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9209–9218, June 2021.

Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

Qifeng Chen and Vladlen Koltun. Photographic image synthesis with cascaded refinement networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 1511–1520, 2017.

Ekin D Cubuk, Barret Zoph, Samuel S Schoenholz, and Quoc V Le. Intriguing properties of adversarial examples. *arXiv preprint arXiv:1711.02846*, 2017.

Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 113–123, 2019.

Josip Djolonga, Jessica Yung, Michael Tschannen, Rob Romijnders, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Matthias Minderer, Alexander D'Amour, Dan Moldovan, et al. On robustness and transferability of convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16458–16468, 2021.

Alexey Dosovitskiy and Thomas Brox. Generating images with perceptual similarity metrics based on deep networks. *Advances in neural information processing systems*, 29:658–666, 2016.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12873–12883, 2021.

Jean-Yves Franceschi, Edouard Delasalles, Mickaël Chen, Sylvain Lamprier, and Patrick Gallinari. Stochastic latent residual video prediction. In *International Conference on Machine Learning*, pp. 3233–3246. PMLR, 2020.

Leon A Gatys, Alexander S Ecker, and Matthias Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015.

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. *arXiv preprint arXiv:1811.12231*, 2018.

Stephen Hanson and Lorien Pratt. Comparing biases for minimal network construction with back-propagation. *Advances in neural information processing systems*, 1:177–185, 1988.

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

Katherine L Hermann, Ting Chen, and Simon Kornblith. The origins and prevalence of texture bias in convolutional neural networks. *arXiv preprint arXiv:1911.09071*, 2019.

Jonathan Huang, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, Alireza Fathi, Ian Fischer, Zbigniew Wojna, Yang Song, Sergio Guadarrama, et al. Speed/accuracy trade-offs for modern convolutional object detectors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7310–7311, 2017.

Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and< 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pp. 694–711. Springer, 2016.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4401–4410, 2019.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8110–8119, 2020.

Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better imagenet models transfer better? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2661–2671, 2019.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

Orest Kupyn, Tetiana Martyniuk, Junru Wu, and Zhangyang Wang. Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8878–8887, 2019.

Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. Learning blind video temporal consistency. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 170–185, 2018.

Alex X Lee, Richard Zhang, Frederik Ebert, Pieter Abbeel, Chelsea Finn, and Sergey Levine. Stochastic adversarial video prediction. *arXiv preprint arXiv:1804.01523*, 2018a.

Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018b.

Yanghao Li, Naiyan Wang, Jiaying Liu, and Xiaodi Hou. Demystifying neural style transfer. *arXiv preprint arXiv:1701.01036*, 2017.

Andrew Liu, Richard Tucker, Varun Jampani, Ameesh Makadia, Noah Snavely, and Angjoo Kanazawa. Infinite nature: Perpetual view generation of natural scenes from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14458–14467, October 2021.

Andreas Lugmayr, Martin Danelljan, Luc Van Gool, and Radu Timofte. Srflow: Learning the super-resolution space with normalizing flow. In *European Conference on Computer Vision*, pp. 715–732. Springer, 2020.

Andreas Lugmayr, Martin Danelljan, and Radu Timofte. Ntire 2021 learning the super-resolution space challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 596–612, June 2021.

Ricardo Martin-Brualla, Noha Radwan, Mehdi S. M. Sajjadi, Jonathan T. Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7210–7219, June 2021.

Alexander Mathis, Thomas Biasi, Steffen Schneider, Mert Yuksekgonul, Byron Rogers, Matthias Bethge, and Mackenzie W. Mathis. Pretraining boosts out-of-domain robustness for pose estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1859–1868, January 2021.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision*, pp. 405–421. Springer, 2020.

John P Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. Accuracy on the line: On the strong correlation between out-of-distribution and in-distribution generalization. In *International Conference on Machine Learning*, pp. 7721–7735. PMLR, 2021.

Reiichiro Nakano. A discussion of 'adversarial examples are not bugs, they are features': Adversarially robust neural style transfer. *Distill*, 2019. doi: 10.23915/distill.00019.4. https://distill.pub/2019/advex-bugs-discussion/response-4.

Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Z Qureshi, and Mehran Ebrahimi. Edgeconnect: Generative image inpainting with adversarial edge learning. *arXiv preprint arXiv:1901.00212*, 2019.

A Emin Orhan. Robustness properties of facebook's resnext wsl models. *arXiv preprint arXiv:1907.07640*, 2019.

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. *arXiv preprint arXiv:2103.00020*, 2021.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pp. 5389–5400. PMLR, 2019.

Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: A stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2287–2296, June 2021.

Gernot Riegler and Vladlen Koltun. Free view synthesis. In *European Conference on Computer Vision*, pp. 623–640. Springer, 2020.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

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*, pp. 407007, 2020.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56): 1929–1958, 2014. URL `http://jmlr.org/papers/v15/srivastava14a.html`.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016a.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016b.

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pp. 6105–6114. PMLR, 2019.

Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *arXiv preprint arXiv:2007.00644*, 2020.

Hugo Touvron, Andrea Vedaldi, Matthijs Douze, and Hervé Jégou. Fixing the train-test resolution discrepancy. *arXiv preprint arXiv:1906.06423*, 2019.

Richard Tucker and Noah Snavely. Single-view view synthesis with multiplane images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 551–560, 2020.

Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P. Srinivasan, Howard Zhou, Jonathan T. Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4690–4699, June 2021.

Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. Synsin: End-to-end view synthesis from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7467–7477, 2020.

Bohan Wu, Suraj Nair, Roberto Martin-Martin, Li Fei-Fei, and Chelsea Finn. Greedy hierarchical variational autoencoders for large-scale video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2318–2328, 2021.

Cihang Xie, Mingxing Tan, Boqing Gong, Alan Yuille, and Quoc V Le. Smooth adversarial training. *arXiv preprint arXiv:2006.14536*, 2020a.

Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10687–10698, 2020b.

Dong Yin, Raphael Gontijo Lopes, Jon Shlens, Ekin Dogus Cubuk, and Justin Gilmer. A fourier perspective on model robustness in computer vision. *Advances in Neural Information Processing Systems*, 32:13276–13286, 2019.

Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, et al. A large-scale study of representation learning with the visual task adaptation benchmark. *arXiv preprint arXiv:1910.04867*, 2019.

Kaihao Zhang, Wenhan Luo, Yiran Zhong, Lin Ma, Bjorn Stenger, Wei Liu, and Hongdong Li. Deblurring by realistic blurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2737–2746, 2020.

Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. Fsim: A feature similarity index for image quality assessment. *IEEE Transactions on Image Processing*, 20(8):2378–2386, 2011. doi: 10.1109/TIP.2011.2109730.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

Youshan Zhang and Brian D. Davison. Impact of imagenet model selection on domain adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, March 2020.

Shengyu Zhao, Jonathan Cui, Yilun Sheng, Yue Dong, Xiao Liang, Eric I Chang, and Yan Xu. Large scale image completion via co-modulated generative adversarial networks. *arXiv preprint arXiv:2103.10428*, 2021.

Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, Hanxiao Liu, Ekin D Cubuk, and Quoc V Le. Rethinking pre-training and self-training. *arXiv preprint arXiv:2006.06882*, 2020.

A Perceptual Scores: High Resolution Models

We train out-of-the-box ImageNet models on their typical resolutions (224 × 224) and above. We then resize the 64 × 64 BAPPS images to the resolutions on which the models are trained on and report their accuracies and PS in Figure 7a. The inverse-U relationship still exists. But as expected, the perceptual scores across all models are considerably worse. Our best and worst PS are 67.6 and 64.2 as compared to 69.1 and 67.3 obtained with models trained directly on 64 × 64 images. Therefore, we advise to train models on smaller images (64 × 64) to have a high PS.

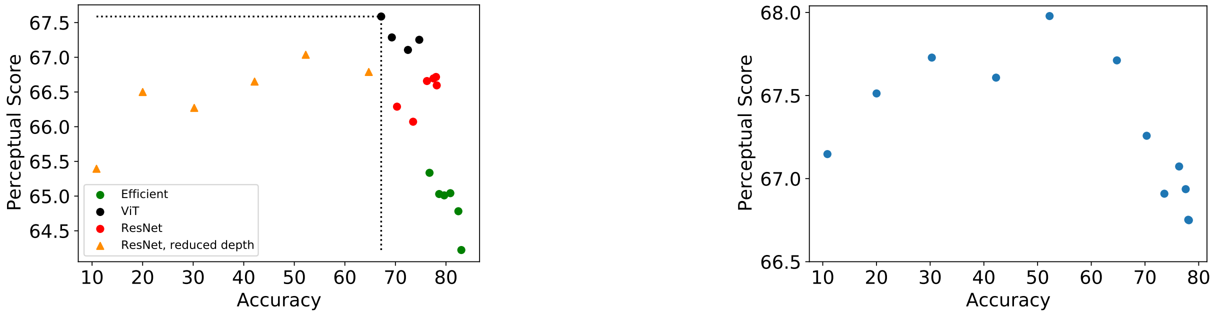
We also evaluate the PS of out-of-the-box ResNets trained on high-resolution images using 64 × 64 BAPPS images directly and observe the same phenomenon in 7b.

B Perceptual Score: Sub-dataset Breakdown

BAPPS consists of two sub-datasets "Distortions" and "Real Algorithms" that differ in the generative process of the target patches. In Figs. 10 and 11, we show scatter plots for accuracy-PS and PS vs hyper-parameter values for the "Distortion" subset. We show similar results for the "Real Algorithms" subset in Figs. 12 and 13. The inverse-U relationship generalizes across each sub-dataset of BAPPS.

C Perceptual Score: Learned

Previous work found that it is possible to improve the prediction of human binary choice on a dataset of bird images Attarian et al. (2020) and human 2-AFC judgements Zhang et al. (2018) by finding appropriate linear transformations. Through our paper is focused on assessing the inherent properties of ImageNet pretrained models to capture perceptual similarity, yet following the linear evaluation protocol in Zhang et al. (2018), we train a linear layer on top of pretrained ImageNet features to match supervised human judgements on BAPPS. The PS gap between ResNet-6 and ResNet-200 narrows down from 1.5 to 0.8, but even after training, there is a negative corelation among high-performing state-of-the-art residual networks in Fig. 8



(a)

(b)

Figure 7: We train all out-of-the-box networks on their typical higher resolutions (224 × 224) and above. In Fig. 7a, we resize the 64 × 64 BAPPS images to match the resolutions at which the corresponding networks are trained. In Fig. 7b, we display the PS evaluated on 64 × 64 images directly on out-of-the-boz ResNets. The inverse-U relationship exists but the perceptual scores across all architectures are considerably worse

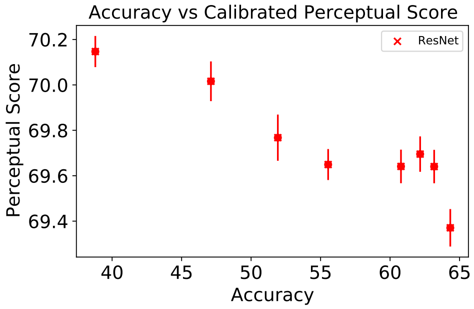


Figure 8: We train a linear layer on top of ImageNet features. The negative correlation exists among state-of-the-art residual networks.

D Perceptual Score: Learning Rate

We demonstrate that the inverse-U phenomenon also exists as a function of the peak learning rate in ResNets in 9.

E Class Granularity: More Results

We replace the “random sampling” strategy in our previous “Class Granularity” experiment with two variants. We compute ResNet-6’s accuracy on each ImageNet class and rank the classes according to the per-class accuracy. In our “top” and “bottom” experiments, each subset of k classes consists of the top k and bottom k classes according to this ranking respectively. In Figs. 14b and 14a, we vary k from 50 to 900. As observed in our previous experiment, as the number of classes are reduced, ResNet-200’s PS improves but still underperforms ResNet-6. Interestingly, the particular sampling strategy used does not seem to have a significant effect.

F Spatial Frequency Analysis: Further Exploration

Normalized Frequency Slope. Figs. 15a and 15b, we display the “normalized accuracy slope” as a function of r. The “normalized accuracy slope” at radius r is the difference between the “normalized accuracy” at radius r and the “normalized accuracy” at radius $r - 1$. It helps us understand: at which frequencies

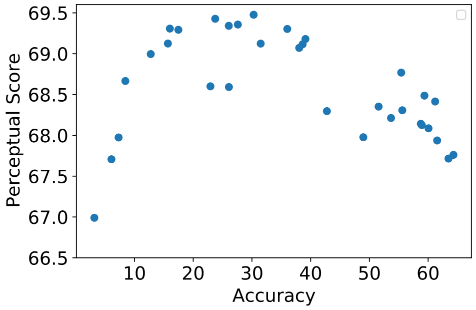
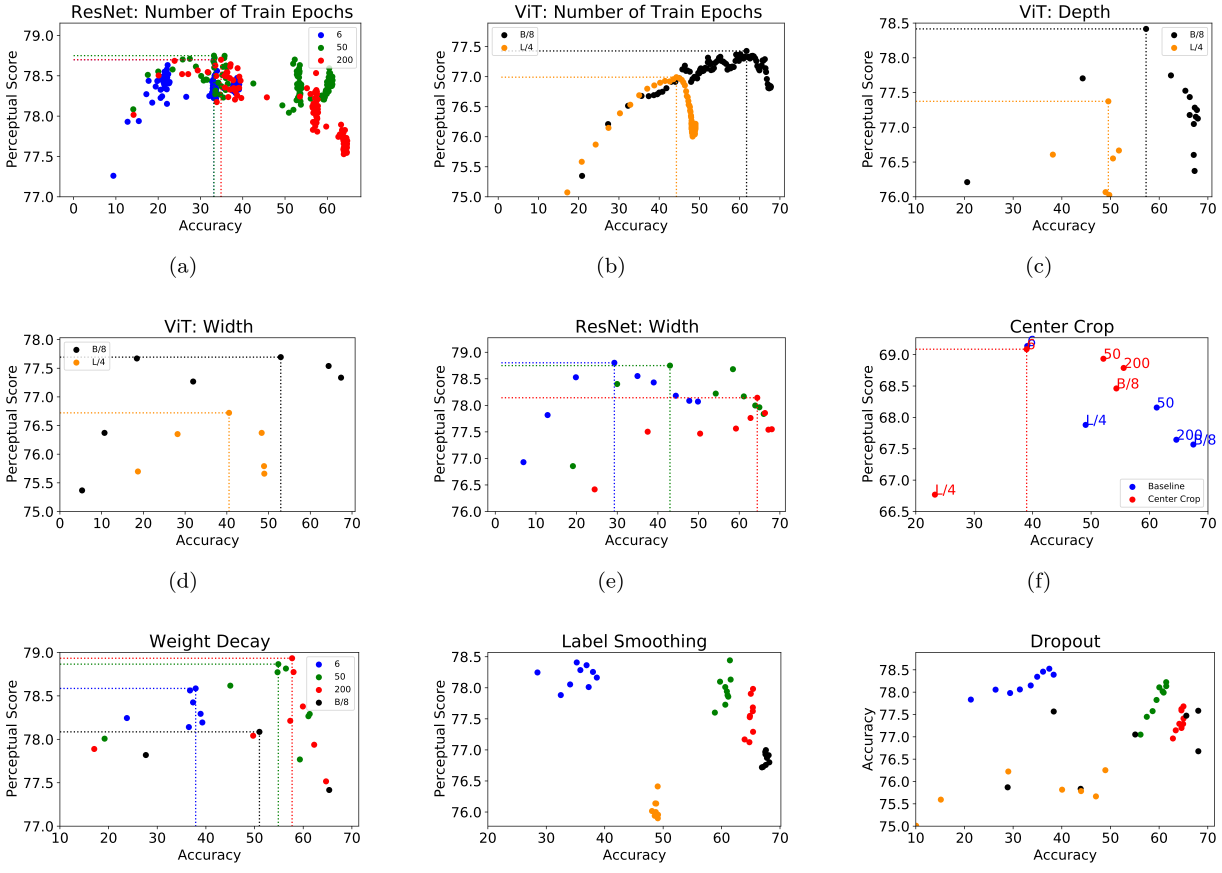


Figure 9: Inverse-U as a function of peak learning rates.



(a)
(b)
(c)

(d)
(e)
(f)

(g)
(h)
(i)

Figure 10: **Distortions:** The relationship between Perceptual Scores and accuracy when varying different hyperparameters on the **tradional algorithms** subset of BAPPS. Each plot depicts the ImageNet accuracy (x axis) and Perceptual Score (y axis) obtained by a 1D sweep over that particular hyperparameter.

do different models have their highest accuracy gains? Models that have high PS such as ResNet-6 and early-stopped ResNet-200 have higher slopes at higher frequencies as compared to low PS models.

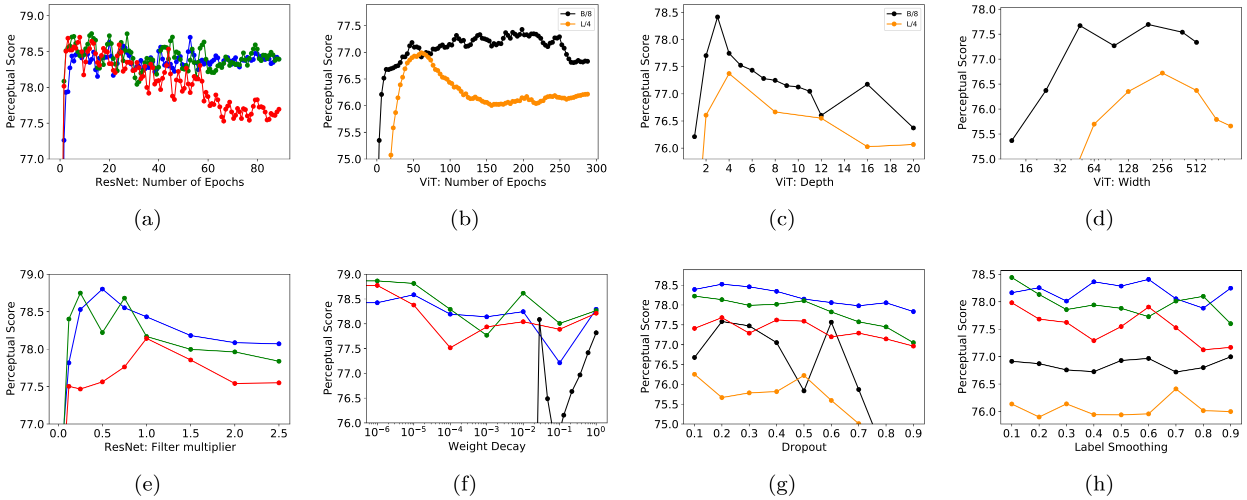


Figure 11: **Distortions:** In each plot, we vary a single hyperparameter along a 1D grid and plot the Perceptual Scores on the **distortions** subset of BAPPS

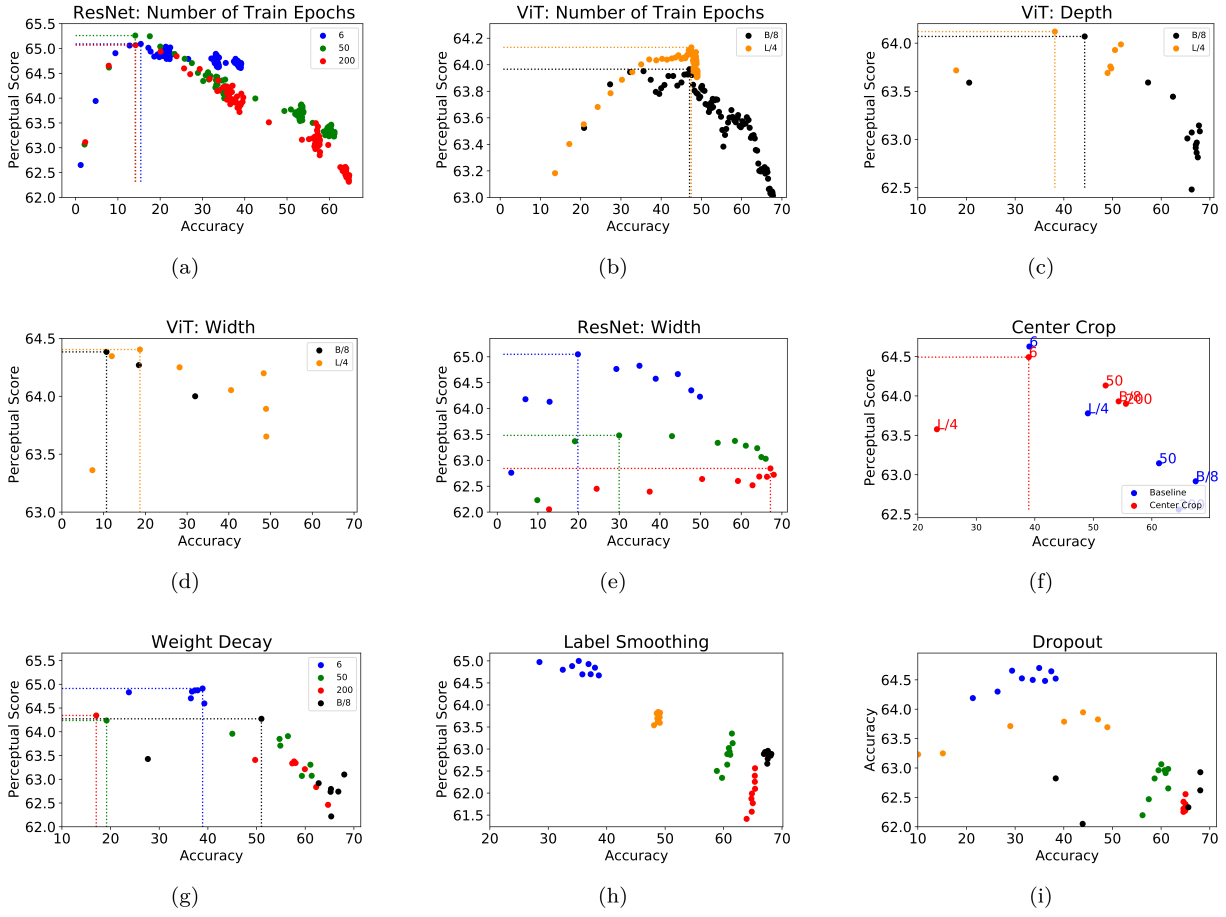
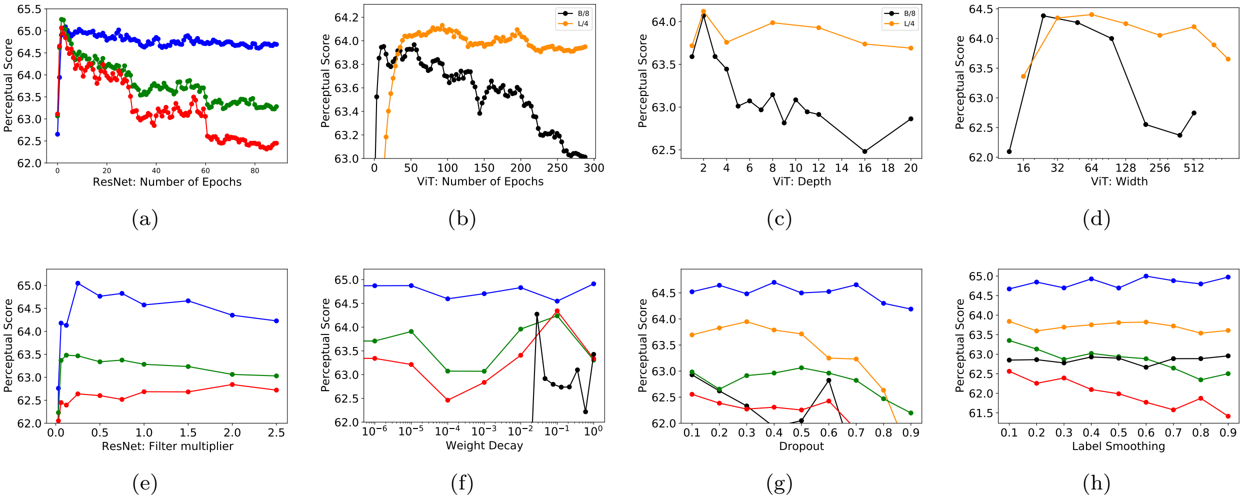


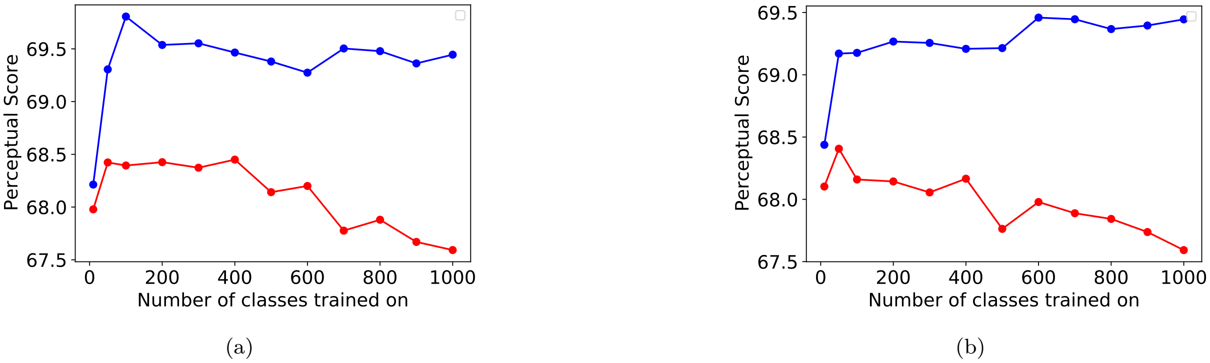
Figure 12: **Real Algorithms:** The relationship between Perceptual Scores and accuracy when varying different hyperparameters on the **real algorithms** subset of BAPPS. Each plot depicts the ImageNet accuracy (x axis) and Perceptual Score (y axis) obtained by a 1D sweep over that particular hyperparameter.



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 13: **Real Algorithms:** In each plot, we vary a single hyperparameter along a 1D grid and plot the Perceptual Scores on the **real algorithms** subset of BAPPS.



(a) (b)

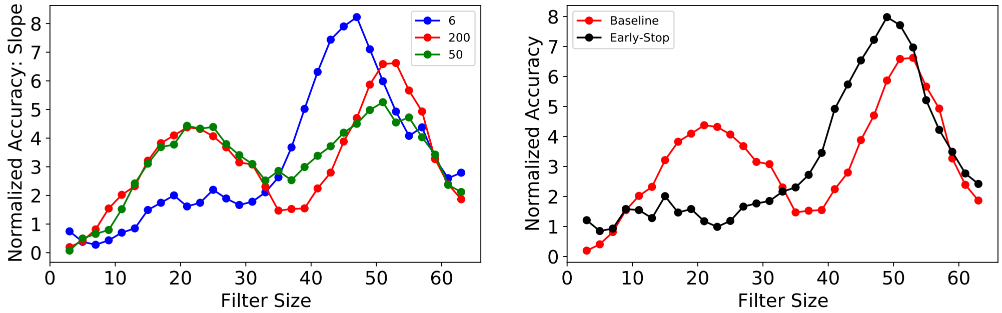
Figure 14: We dive deep into the impact of class granularity. We create class subsets according to ResNet-6's accuracy, instead of a random subset. In Fig. 14a and Fig. 14b, we show the accuracy and PS of the networks trained on the top and botton x number of classes as given by ResNet-6's accuracy.

G Augmentations:

We investigate the effect of Gaussian Noise Augmentation and AutoAugment on PS. By construction, any augmentation strategy forces the network to produce the same outputs at the last layer of the network, for different distortions of an image. One can thus expect augmentation to make a network less sensitive to various distortions of an image, and as a result lower its PS.

Gaussian Noise. For a given noise factor s, we first sample $\sigma \sim U(0, s)$ per-image. We then sample per-pixel noise independently from the Gaussian distribution $\mathcal{N}(0, \sigma)$, truncate it between 0.0 and 1.0 and add it to the original image. We show results for 4 noise factors, 0.1, 0.2, 0.5 and 1.0. The "Distortions" subset of BAPPS consists of Gaussian-noise based distortions, and Gaussian noise augmentation would make the network less-sensitive to such distortions. As an effect, in Fig. 16b, an the noise factor increases, the PS on the "Distortions" subset of BAPPS decreases by a huge amount.

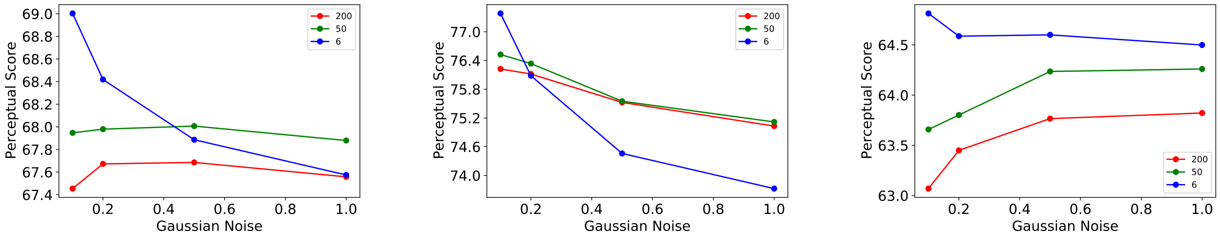
AutoAugment. We train models with a state-of-the-art augmentation technique AutoAugment Cubuk et al. (2019) that consists of a mixture of different augmentations. As seen in Figs. 16d, 16f, 16e, for every



(a)

(b)

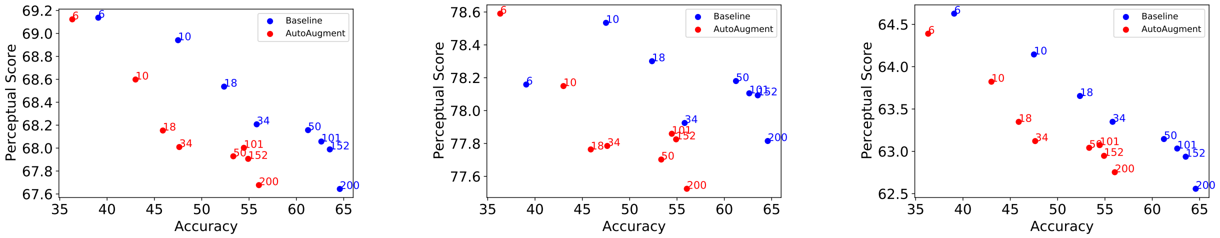
Figure 15: We further explore the spatial frequency responses to networks via **Normalized Frequency Slopes:** (Figs. 15a and 15b)



(a) All

(b) Traditional

(c) Real

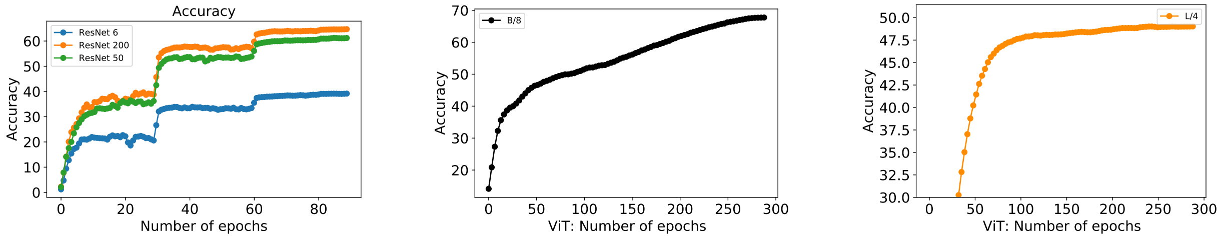


(d) All

(e) Traditional

(f) Real

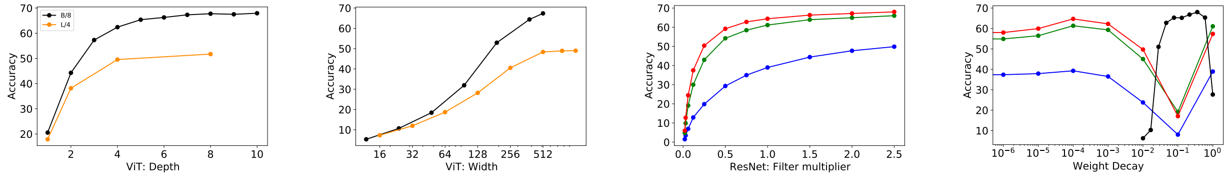
Figure 16: We investigate the effect of Gaussian Noise Augmentation and AutoAugment strategies on PS



(a)

(b)

(c) Real



(d)

(e)

(f)

(g)

Figure 17: In each plot, we vary a single hyperparameter along a 1D grid and plot the validation accuracy.

architecture, AutoAugment decreases both the accuracy and PS. We note that AutoAugment strategy was developed explicitly by maximizing the validation accuracy on high resolution (224 × 224) images, and thus the policy might not transfer to low resolution (64 × 64) images. This can explain the decrease in accuracy

H Accuracy vs Hyperparameters

We plot the accuracy as a function of each hyperparameter where we observe the inverse-U relationshop between validation accuracy and Perceptual Score in Fig. 17. We note that the validation accuracy steadily increases as a function of the hyperparameter. The peak in Perceptual Scores occur in the underfitting regime where the ImageNet models have poor to moderate accuracies.

I Code Snippets

We present code snippets for the different distance functions used in our paper.

Listing 1: Code Snippets for different perceptual functions

```
def perceptual(tensor1, tensor2, eps=1e-10):
  """Default perceptual distance function.

  Args:
    tensor1: shape=(B, H, W, C)
    tensor2: shape=(B, H, W, C)
  Returns:
    dist: shape=(B,)
  """
  tensor1_n = np.linalg.norm(tensor1, ord=2, axis=-1, keepdims=True)
  tensor2_n = np.linalg.norm(tensor2, ord=2, axis=-1, keepdims=True)
  tensor1 = tensor1 / (tensor1_n + eps)
  tensor2 = tensor2 / (tensor2_n + eps)
  dist = np.sum((tensor1 - tensor2)**2, axis=-1)
  dist = np.mean(dist, axis=(1, 2))
  return dist
```

```
def mean_pool(tensor1, tensor2, eps=1e-10):
  """Mean Pool perceptual distance function.

  Args:
    tensor1: shape=(B, H, W, C)
    tensor2: shape=(B, H, W, C)
  Returns:
    dist: shape=(B,)
  """
  tensor1 = np.mean(tensor1, (1, 2), keepdims=True)
  tensor2 = np.mean(tensor2, (1, 2), keepdims=True)
  return perceptual(tensor1, tensor2, eps=eps)


def compute_gram(tensor, eps):
  """Returns (BxCxC) cross correlation."""
  _, h, w, c = tensor.shape
  tensor = np.reshape(tensor, (-1, h*w, c))
  tensor_norm = np.linalg.norm(tensor, ord=2, axis=1, keepdims=True)
  tensor = tensor / (tensor_norm + eps)
  # Channel-wise cross correlation.
  tensor_t = np.transpose(tensor, (0, 2, 1))
  return np.matmul(tensor_t, tensor)


def style(tensor1, tensor2, eps=1e-10):
  """Style perceptual distance function.

  Args:
    tensor1: shape=(B, H, W, C)
    tensor2: shape=(B, H, W, C)
  Returns:
    dist: shape=(B,)
  """
  tensor1_gram = compute_gram(tensor1, eps=eps)
  tensor2_gram = compute_gram(tensor2, eps=eps)
  dist = tensor1_gram - tensor2_gram
  return np.mean(dist**2, axis=(1, 2))
```

J Default Hyper-parameters

We provide the default training hyper-parameters for the ResNets, EfficientNets and Vision Transformers in Tables 2, 3, 4 and 5.

Hyper-Parameter	Value
Batch Size	1024
Base Learning Rate	0.1
Train Steps	112590
Momentum	0.9
Weight Decay	0.0001
Label Smoothing	0.0
LR Schedule	Step-wise Decay
Batch-Norm Momentum	0.9

Table 2: ResNet: Default Hyperparameters

Hyper-Parameter	Value
Batch Size	2048
Base Learning Rate	0.128
Train Steps	218949
Optimizer	RMSProp
Momentum	0.9
Weight Decay	1e-5
Label Smoothing	0.1
LR Schedule	Warmup + Exp Decay
Batch-Norm Momentum	0.9
Polyak Average	0.9999

Table 3: EfficientNet: Default Hyperparameters

Hyper-Parameter	Value
Batch Size	4096
Base Learning Rate	3e-3
LR Schedule	Warmup + Cosine
LR Warmup Steps	10000
Train Steps	93834
Optimizer	Adam
Beta1	0.9
Beta2	0.999
Weight Decay	0.3
Hidden Size	768
MLP Dim	3072
Number of Layers	12
Number of Heads	12
Dropout	0.1
Label Smoothing	0.1

Table 4: ViT-B/8: Default Hyperparameters

Hyper-Parameter	Value
Batch Size	4096
Base Learning Rate	1e-4
LR Schedule	Warmup + Cosine
LR Warmup Steps	10000
Train Steps	93834
Optimizer	Adam
Beta1	0.9
Beta2	0.999
Weight Decay	0.01
Hidden Size	1024
MLP Dim	4096
Number of Layers	24
Number of Heads	12
Dropout	0.1
Label Smoothing	0.1

Table 5: ViT-L/4: Default Hyperparameters