ConvNet vs Transformer, Supervised vs CLIP: Beyond ImageNet Accuracy

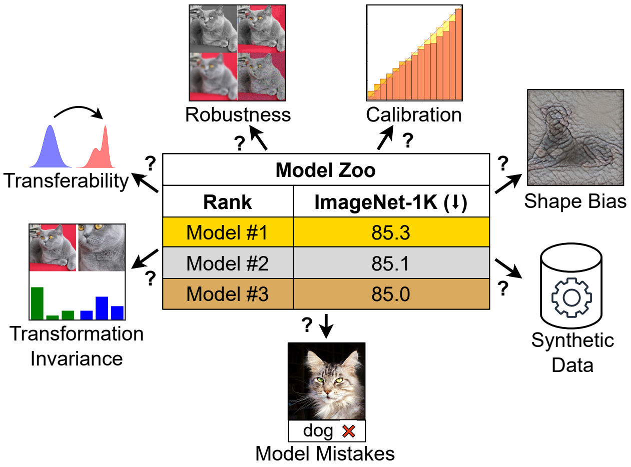
Kirill Vishniakov [1] Zhiqiang Shen [1] Zhuang Liu [2]

Abstract

Modern computer vision offers a great variety of models to practitioners, and selecting a model from multiple options for specific applications can be challenging. Conventionally, competing model architectures and training protocols are compared by their classification accuracy on ImageNet. However, this single metric does not fully capture performance nuances critical for specialized tasks. In this work, we conduct an in-depth comparative analysis of model behaviors beyond ImageNet accuracy, for both ConvNet and Vision Transformer architectures, each across supervised and CLIP training paradigms. Although our selected models have similar ImageNet accuracies and compute requirements, we find that they differ in many other aspects: types of mistakes, output calibration, transferability, and feature invariance, among others. This diversity in model characteristics, not captured by traditional metrics, highlights the need for more nuanced analysis when choosing among different models. Code is available at github.com/kirill-vish/Beyond-INet.

1. Introduction

The computer vision model landscape has become increasingly complex. From early ConvNets (LeCun et al., 1998) to advances in Vision Transformers (Dosovitskiy et al., 2020), the variety of models available has expanded significantly. Similarly, training paradigms have evolved from supervised training on ImageNet (Deng et al., 2009) to self-supervised learning (Chen et al., 2020; He et al., 2020) and image-text pair training like CLIP (Radford et al., 2021). While signaling progress, this explosion of choices poses a significant challenge for practitioners: how to select a model that suits their purposes?



Model Zoo	
Rank	ImageNet-1K (↓)
Model #1	85.3
Model #2	85.1
Model #3	85.0

Figure 1. Models are often compared only by their ImageNet accuracy, without looking at many other important behaviors. In our work, we analyze models with similar ImageNet accuracies and find that they have vastly different properties.

Conventionally, ImageNet accuracy has served as the primary metric for evaluating model performance. It has driven remarkable progress since it ignited the deep learning revolution (Krizhevsky et al., 2012). However, this metric is becoming increasingly insufficient. While ImageNet is useful to measure a model's general capability, it does not capture the nuanced differences arising from varying architectures, training paradigms, and data – models with different properties may appear similar if judged solely based on ImageNet accuracy (Fig. 1). This limitation becomes more pronounced as models start to overfit the idiosyncrasies of ImageNet with saturated accuracies (Beyer et al., 2020).

A particularly noteworthy example is CLIP. Despite having a similar ImageNet accuracy as a ResNet (He et al., 2016), CLIP's vision encoder exhibits much better robustness and transferability. This has sparked research that explores and builds upon the unique strengths of CLIP (Ramesh et al., 2022; Luo et al., 2022; Wortsman et al., 2022; Vinker et al., 2023), which were not evident from the ImageNet metric alone. This demonstrates that analyzing alternative properties could help discover useful models.

In addition to fundamental research, the growing integration of vision models into production systems also calls for a deep understanding of their behaviors. Conventional metrics do not fully capture models' ability to handle real-world vision challenges like varying camera poses, lighting condi-

[1] MBZUAI [2] Meta AI Research. Correspondence to: Kirill Vishniakov <ki.vishniakov@gmail.com>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

Model	Architecture	Pretraining	Finetuning	Paradigm	FLOPs	#Param	INet-1K val%
ViT-sup	ViT-B/16	ImageNet-21K	ImageNet-1K	supervised	17.5G	87M	85.5
ConvNeXt-sup	ConvNeXt-B	ImageNet-21K	ImageNet-1K	supervised	15.4G	89M	85.5
ViT-clip	ViT-B/16	LAION-400M	—	CLIP	17.5G	87M	67.0
ConvNeXt-clip	ConvNeXt-B	LAION-400M	—	CLIP	15.4G	89M	66.3

Table 1. **Model summary in our analysis.** We select ConvNeXt and ViT with similar ImageNet accuracies within each training paradigm.

tions, or occlusions. For instance, models trained on datasets such as ImageNet often struggle (Yamada & Otani, 2022) to transfer their performance to real-world applications where conditions and scenarios are more diverse.

To bridge this gap, we conduct an in-depth exploration focusing on model behaviors beyond ImageNet accuracy. We analyze four leading models in the computer vision: ConvNeXt (Liu et al., 2022), as a representative ConvNet, and Vision Transformer (ViT) (Dosovitskiy et al., 2020), each under supervised and CLIP training. The selected models are similar in parameter counts and show nearly identical accuracy on ImageNet-1K within each training paradigm, ensuring a fair comparison. Our study delves into a wide array of model characteristics, such as types of prediction errors, generalization capabilities, invariances of the learned representations, calibration, and many others. Importantly, our focus is on properties exhibited by the model without additional training or finetuning, providing insights for practitioners interested in using pretrained models directly.

In our analysis, we discover substantial variations in model behaviors among different architectures and training paradigms. For example, CLIP models make fewer classification errors relative to their ImageNet performance. However, supervised models are better calibrated and superior on ImageNet robustness benchmarks. ConvNeXt has an advantage on synthetic data but is more texture-biased than ViT. We also find that supervised ConvNeXt excels on many benchmarks and achieves transferability comparable to that of CLIP models. Based on these findings, it becomes evident that various models demonstrate their strengths in unique ways that are not captured by a single metric. Our research emphasizes the need for more detailed evaluation metrics for accurate, context-specific model selection and the creation of new benchmarks unrelated to ImageNet.

2. Models

For analyzing ConvNets and Transformers, many previous works (Naseer et al., 2021; Minderer et al., 2021; Zhou et al., 2021; Bai et al., 2021) compare ResNet and ViT. This comparison is often disadvantageous for ConvNet since ViTs are typically trained with more advanced recipes, achieving higher ImageNet accuracy. ViT also has architecture design elements, e.g., LayerNorm (Ba et al., 2016), that were not incorporated in ResNet when it was invented years ago. For a more balanced evaluation, we compare ViT with ConvNeXt (Liu et al., 2022), a modern of ConvNet that matches Transformers' performance.

As for the training paradigms, we compare supervised and CLIP. Supervised models continue to show state-of-the-art performance in computer vision (Dehghani et al., 2023). CLIP models, on the other hand, excel in generalization and transferability, and offer intriguing representational properties that connect vision and language. Self-supervised models (He et al., 2022; Woo et al., 2023) are not included in the results as they showed behaviors similar to supervised models in our preliminary tests. This could be due to their final ImageNet-1K supervised finetuning, which is necessary for studying many properties.

The selected models have similar ImageNet-1K validation accuracies within their respective training paradigms, ensuring a fair comparison. For CLIP models, these indicate their zero-shot accuracies. The models also have similar sizes and computational requirements, and are publicly available. Since we are using pretrained models, we cannot control for the number and quality of data samples seen during training.

For supervised models, we use a pretrained DeiT3-Base/16 (Touvron et al., 2022) for ViT, which shares the same architecture as ViT-Base/16 with an improved training recipe, and ConvNeXt-Base (Liu et al., 2022). For CLIP models, we use vision encoders of ViT-Base/16 and ConvNeXt-Base from OpenCLIP (Ilharco et al., 2021). Note that these models have a slightly different performance from the original OpenAI models (Radford et al., 2021). A detailed model comparison is given in Table 1.

We recognize there are other ViT CLIP models pretrained on larger datasets such as LAION-2B, DataComp (Gadre et al., 2023), DFN (Fang et al., 2023a), which show a better performance. However, OpenCLIP offers only a few pretrained ConvNeXt models and for most of them there is no matching ViT counterpart in terms of ImageNet accuracy, pretraining dataset and parameter count. Therefore, we chose CLIP models pretrained on LAION-400M, offering the most fair comparison between ConvNet and ViT.

Additional models of different sizes. In the Appendix A, we present the performance results for models of different sizes to evaluate the impact of model size on performance. These models are assessed across several benchmarks, including ImageNet-X, PUG-ImageNet, calibration, invariance, and shape/texture bias.

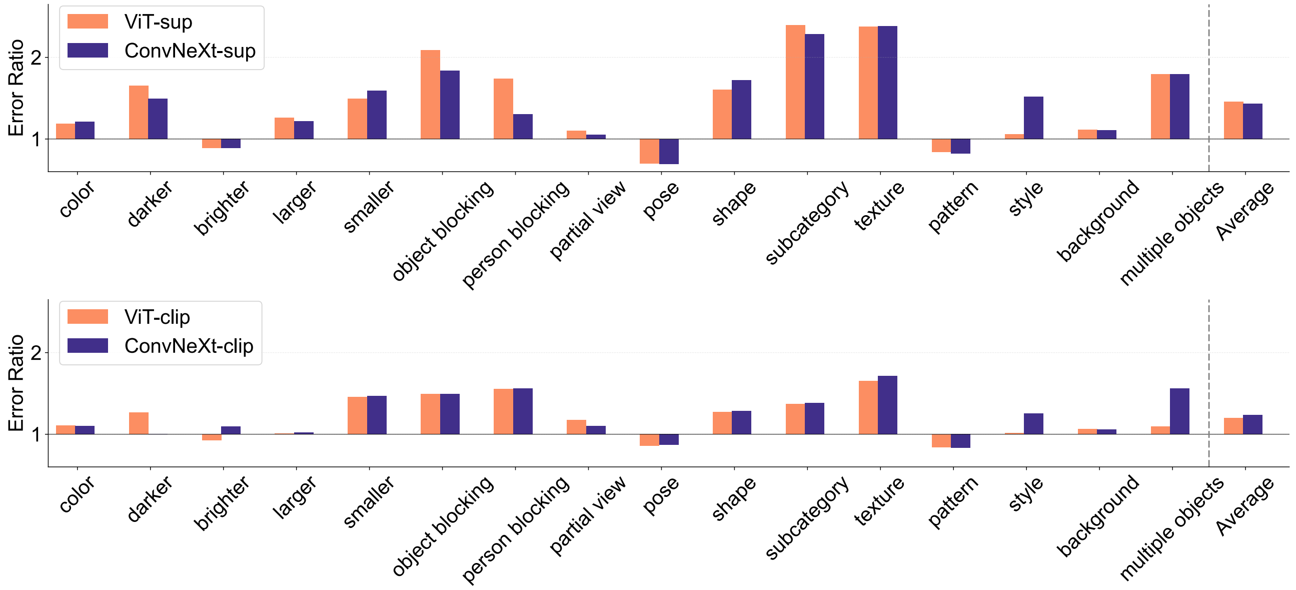


Figure 2. **Model mistakes on ImageNet-X.** Lower is better. ConvNeXt and ViT perform similarly within each training category. CLIP models achieve lower error ratios compared to supervised.

3. Property Analysis

Our analysis is designed to investigate model behaviors that can be evaluated without the need for further training or finetuning. This approach is particularly relevant for practitioners with limited compute resources, who often rely on pretrained models. While we recognize the value of downstream tasks like object detection, our focus is on properties that offer insights with minimal computational demands and reflect behaviors important for real-world applications.

3.1. Model Mistakes

In image classification, a model mistake is an incorrect label assignment, such as misclassifying a cat as a dog. Simply identifying mistaken object classes might not offer actionable insights for model improvement. The key aspect, therefore, is finding the specific reasons for these mistakes. For instance, some models may be particularly sensitive to certain aspects of the data distribution, like texture variations. In this case, a model might consistently make mistakes when the texture of the object differs from what it has been trained on. Identifying mistake types allows for targeted data collection and retraining, improving over a black-box approach.

The ImageNet-X dataset (Idrissi et al., 2022) offers detailed human annotations for 16 factors of variation, such as pose, style, and others. This allows a focused analysis of models' mistake types. The annotations enable measuring model error ratios for each factor independently: $\text{error ratio(factor)} = \frac{1-\text{accuracy(factor)}}{1-\text{accuracy(overall)}}$, where accuracy(overall) is the overall ImageNet-1K validation accuracy, and accuracy(factor) is the accuracy on all the images where the factor was highlighted. This metric measures the model performance on a given factor relative to its overall performance. Lower error ratios indicate better performance, implying higher accuracy for the specific factor. Our results on ImageNet-X are presented in Fig. 2.

CLIP models make fewer mistakes relative to their ImageNet accuracy than supervised. The diagram in Fig. 2 shows that CLIP models have a smaller error ratio, indicating a significant advantage over supervised models. However, it is important to note that the error ratio is relative to overall ImageNet accuracy, where a significant 18% gap exists between supervised and CLIP zero-shot models. In particular, CLIP models are much more robust towards shape, subcategory, texture, object blocking, and darker factors. The key reason for the success of CLIP models is likely the more diverse data used for training.

All models suffer mostly from complex factors like occlusion. For CLIP models, there are three factors with dissimilar performance between ConvNeXt and ViT: multiple objects, style, and darker. For the first two, the ConvNeXt has a higher error ratio, while for the latter, it has an advantage over ViT. For supervised models, the performance only diverges for style and person blocking. Except for these factors, models largely have similar error ratios. The six factors for which all the models have a high error ratio are smaller, object blocking, person blocking, shape, subcategory, and texture. High error ratio factors usually involve complex visual scenarios, which helps to explain why models often make mistakes in these situations. For example, in occlusion, the model often misclassifies due to focusing on the visible, obscuring object.

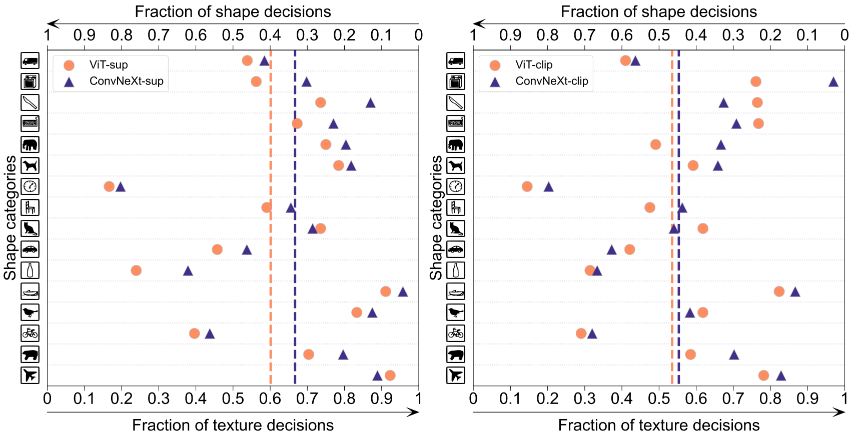


Figure 3. **Fraction of shape vs texture decisions on cue-conflict dataset.** ViT models show a higher shape bias. CLIP models are less texture-biased than their supervised counterparts. All models still have a significant fraction of texture decisions.

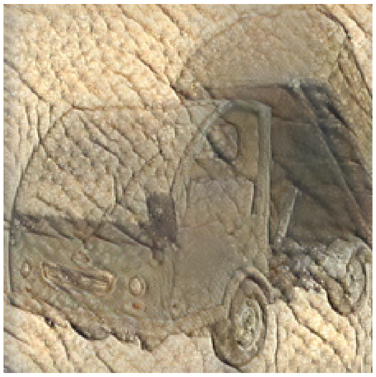


Figure 4. A cue-conflict image (Geirhos et al., 2018).

Texture is the most challenging factor for all models. Interestingly, all models in our analysis have the largest error ratio on the texture factor. It refers to images where the texture of the object differs from its standard appearance. This suggests that models of the current generation largely suffer because of texture bias.

3.2. Shape / Texture Bias

In contrast to humans, who generally use high-level visual cues for recognition, neural networks often rely on more brittle shortcut features (Geirhos et al., 2020). The study of shape-texture bias (Geirhos et al., 2018) serves to highlight this phenomenon by examining model behavior on cue-conflict images, which contain a shape from one class superimposed with the texture from another (Fig. 4). Two key metrics are introduced to quantify this bias: the shape and the texture fractions. The shape fraction calculates the proportion of decisions leaning towards the class represented by the shape, while the texture fraction measures those for the texture class. These metrics reveal whether the classifier favors shape or texture when they conflict.

The study in (Geirhos et al., 2018) showed that ConvNets have a strong bias towards texture, as opposed to shape, which differs from humans. Subsequent work (Naseer et al., 2021) concluded that ViT is less biased towards the texture than ConvNet by comparing the first generation of DeiT-S (Touvron et al., 2021) and ResNet-50. Remarkably, scaling large Transformer models has led to shape biases comparable to human level (Dehghani et al., 2023).

We evaluate shape-texture bias in our models using cue-conflict images and display the findings in Fig. 3. Dashed lines represent average shape bias aggregated over all the categories. Individual markers on horizontal lines depict shape bias for the particular class, which is identified by a corresponding logo on the y-axis. The shape fraction is represented on the top x-axis of the diagrams, while the bottom x-axis indicates the texture fraction.

CLIP models have smaller texture bias than supervised. In Fig. 3, we can observe that ViTs exhibit stronger shape bias than ConvNeXts for both supervised and CLIP models. This is possibly because ConvNeXt is more inclined to learn local features related to textures due to the local nature of convolution operation. However, the gap between ViT and ConvNeXt is much smaller for CLIP-based models. Notably, the shape bias in CLIP models improved by 7% and 12% for both architectures, prompting questions about the benefits of further scaling the training data. ConvNets typically exhibit lower shape bias compared to ViT, however, the gap for CLIP models is marginal. In (Dehghani et al., 2023), it has been shown that a 22B parameter ViT model can achieve 87% shape bias. In our analysis, the ViT CLIP model achieved a maximum shape bias of 46.4%, suggesting that the model size might also play an important role.

3.3. Model Calibration

Besides vulnerability to shortcut features, poor model performance can often be attributed to miscalibration, where a model's confidence in its predictions does not align with actual accuracy. Model calibration is a metric that quantifies the reliability of a model's predicted confidence levels (Guo et al., 2017). A model's confidence for a prediction is defined as the max probability among all classes in its output distribution. We are interested in determining whether the model is overly confident or too uncertain in its predictions. For instance, if the network deems a set of predictions to be 80% confident, does the actual accuracy hover around 80%?

The calibration rate can be quantified by Expected Calibration Error (ECE). To calculate ECE, predictions first need

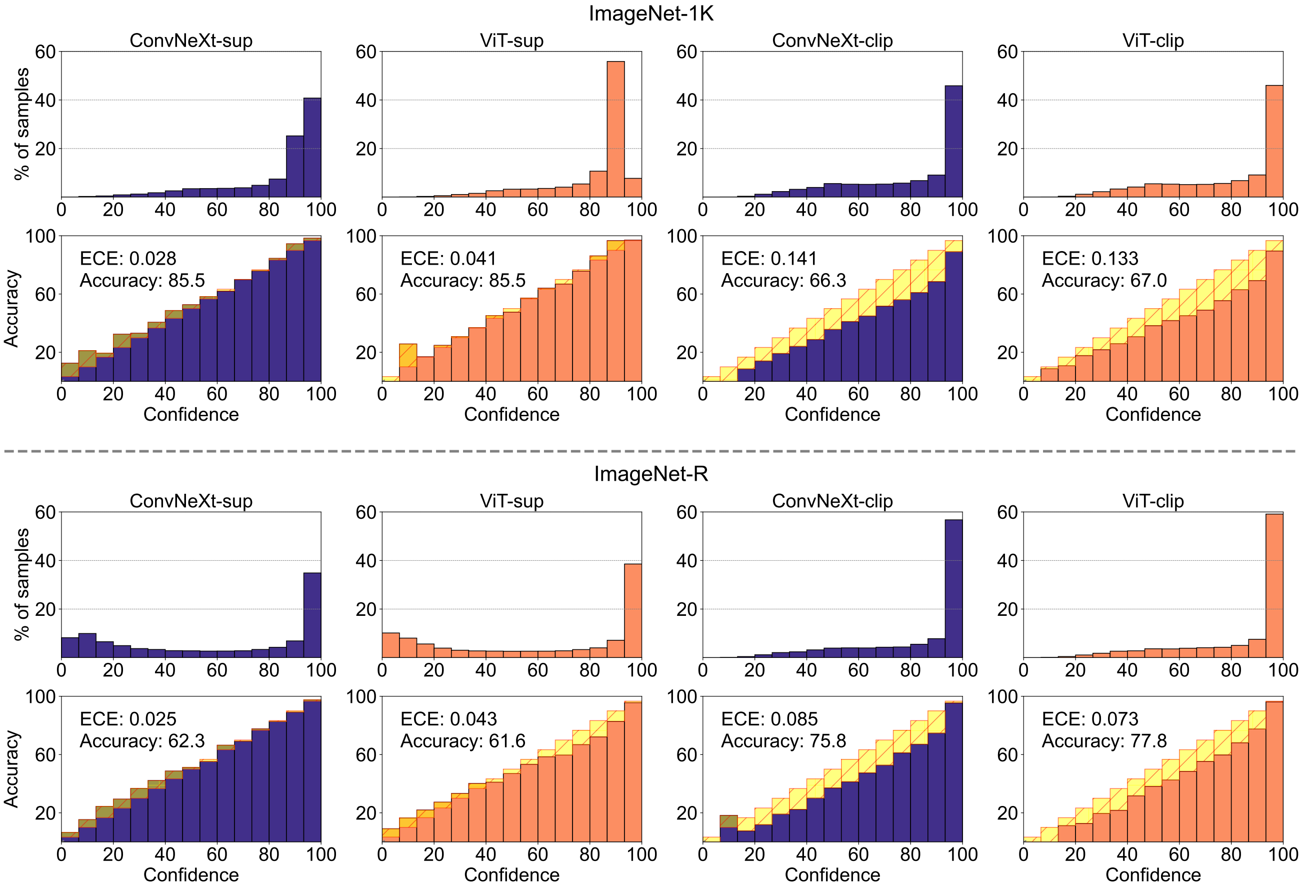


Figure 5. **Calibration results**: confidence histograms (1 and 3 row), reliability diagrams (2 and 4 row), and ECE on ImageNet-1K (top) and ImageNet-R (bottom). Supervised models have lower ECE in both cases. CLIP models have bars under diagonal and many high confidence predictions, indicating overconfidence. ConvNeXt is better (ImageNet-1K) or competitive (ImageNet-R) to ViT.

to be separated into the M bins $B_1, \dots, B_M$ based on their confidence. For instance, one bin can include all the predictions with confidence between 50% and 60% and so on. Each bin's confidence and accuracy are calculated as the average confidence and accuracy of predictions in B_i, represented as $\mathrm{conf}(B_i)$ and $\mathrm{acc}(B_i)$. Then, ECE can be defined as: $\mathrm{ECE} = \sum_i^M \frac{|B_i|}{n} \left|\mathrm{acc}(B_i) - \mathrm{conf}(B_i)\right|$, where $|B_i|$ is the size of the i-th bin.

Model calibration is also often assessed through visualizations, including reliability diagrams and confidence histograms. Reliability diagrams plot the predicted confidence against accuracy; a well-calibrated model would show a graph where points closely align with the diagonal. Confidence histograms display how often different confidence levels occur in the model's predictions.

For a balanced evaluation, we present calibration metrics on two different datasets: ImageNet-1K for in-distribution data and ImageNet-R (Hendrycks et al., 2021a) for out-of-distribution data. We select ImageNet-R as the out-of-distribution because CLIP models show higher accuracy on it than supervised. In all experiments, we use $M = 15$ bins. We plot confidence histograms (1 and 3 rows), reliability diagrams (2 and 4 rows), and ECE in Fig. 5.

CLIP models are overconfident and supervised models are slightly underconfident. In Fig. 5, we observe that CLIP models have bars consistently below the diagonal in reliability diagrams and a notably high last bar in the confidence histogram, signaling overconfidence in both in-distribution and out-of-distribution data. Although (Minderer et al., 2021) attributes calibration performance mainly to architecture, our results suggest otherwise: higher ECE scores in CLIP models, despite superior accuracy on ImageNet-R, indicate that training data and objectives could be more influential factors. We also highlight that our results are different from (Minderer et al., 2021) for CLIP models presumably because they use checkpoints from OpenAI (Radford et al., 2021) and we use from OpenCLIP (Ilharco et al., 2021). In the lower part of Fig. 5 related to ImageNet-R, we note that supervised models exhibit a higher density in the lower confidence intervals of the confidence histograms (3 row). Additionally, these models show elevated accuracy levels in the initial bins of the reliability diagrams (4 row). These findings suggest that supervised models tend to be slightly underconfident on ImageNet-R.

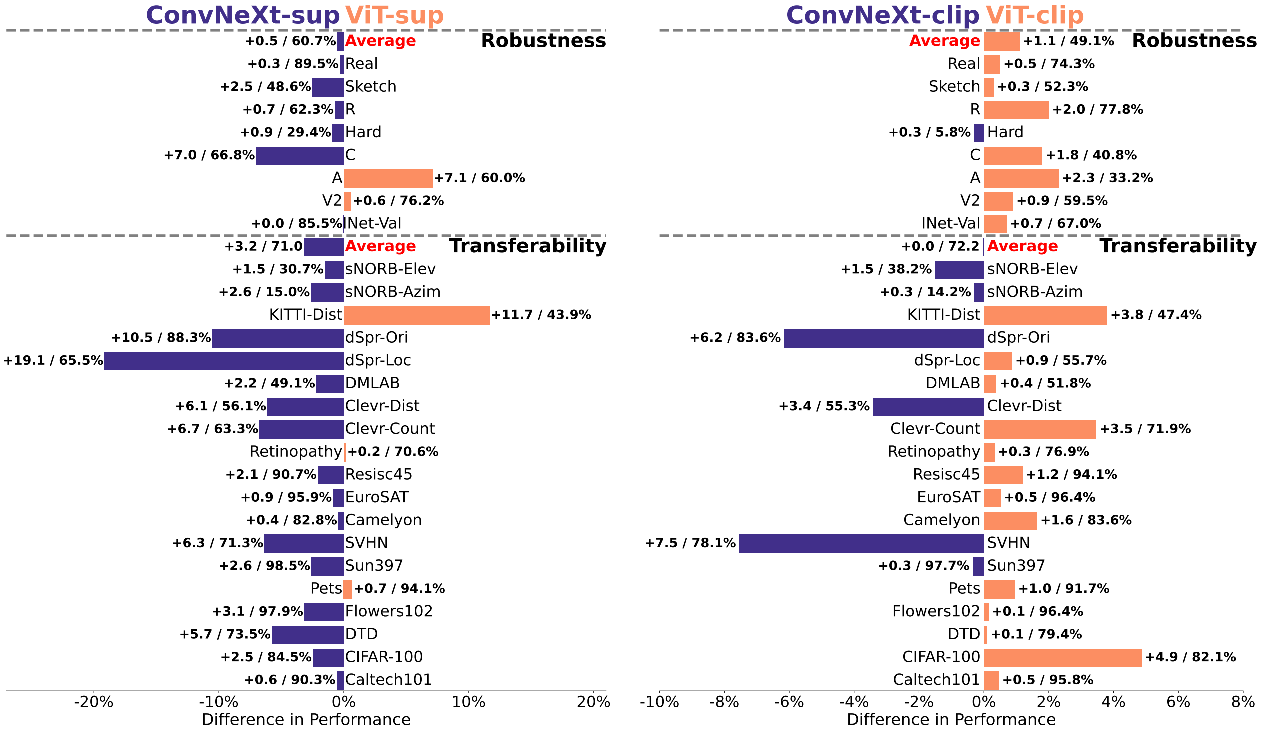


Figure 6. **Robustness (top) and transferability (bottom) results.** CLIP models excel in transferability, while supervised models are better on robustness benchmarks. In transferability, supervised ConvNeXt outperforms supervised ViT and is close to CLIP models.

Supervised ConvNeXt is better calibrated than supervised ViT. Our experiments reveal that supervised ConvNeXt outperforms Transformer in calibration, contrary to (Minderer et al., 2021)'s findings on ViTs and ConvNets. This discrepancy is because (Minderer et al., 2021) focused on older ConvNet architectures, such as ResNet, while we use a more modern one. For CLIP models, we find that ViT is only slightly better than ConvNeXt.

3.4. Robustness

A model may excel on data from its training distribution but struggle to generalize to a distribution shift (Recht et al., 2019). These shifts can arise from natural perturbations such as atmospheric conditions (e.g., fog, rain), camera noise, or variations in object location and orientation. Model robustness quantifies a model's capability to adapt to changes in data distributions. A robust model should maintain high accuracy with these perturbations. This is particularly important for applications where reliability is a primary concern.

We evaluate the robustness on several ImageNet variants that feature many types of natural variations and corruptions: V2 (Recht et al., 2019), A (Hendrycks et al., 2021b), C (Hendrycks & Dietterich, 2019), R (Hendrycks et al., 2021a), Sketch (Wang et al., 2019), Real (Beyer et al., 2020), and Hard (Taesiri et al.). We also provide ImageNet-1K validation accuracy for reference (INet-Val). The results are shown in Fig. 6 (top).

Supervised models are better than CLIP on most of the robustness benchmarks. In Fig. 6, we can see that supervised models perform better than CLIP on most datasets except ImageNet-R and ImageNet-Sketch. CLIP models' success on ImageNet-R and ImageNet-Sketch suggests they handle abstract or creative visuals better than supervised models. The advantage of supervised models is likely related to the fact that all robustness datasets share the same set of classes as the original ImageNet-1K, on which they were finetuned. This underscores the need for the development of new robustness benchmarks that are not directly related to ImageNet. Additionally, CLIP models may achieve higher performance when pretrained on larger datasets (Fang et al., 2023a; Gadre et al., 2023). ViT and ConvNeXt, on average, have similar performance across both supervised and CLIP.

3.5. Transferability

The transfer learning performance of a model indicates its ability to adapt to new tasks and datasets beyond its original training domain (Kolesnikov et al., 2020). Good transferability allows for rapid finetuning with minimal additional effort, making it easier to scale the model to a wide range of real-world applications. The ability of a model to adapt to these shifts without significant degradation in performance serves as a valuable metric for its utility and generalization capabilities. For instance, consider a model that has been originally trained on ImageNet, which primarily consists of natural images. A test of its transferability would be to

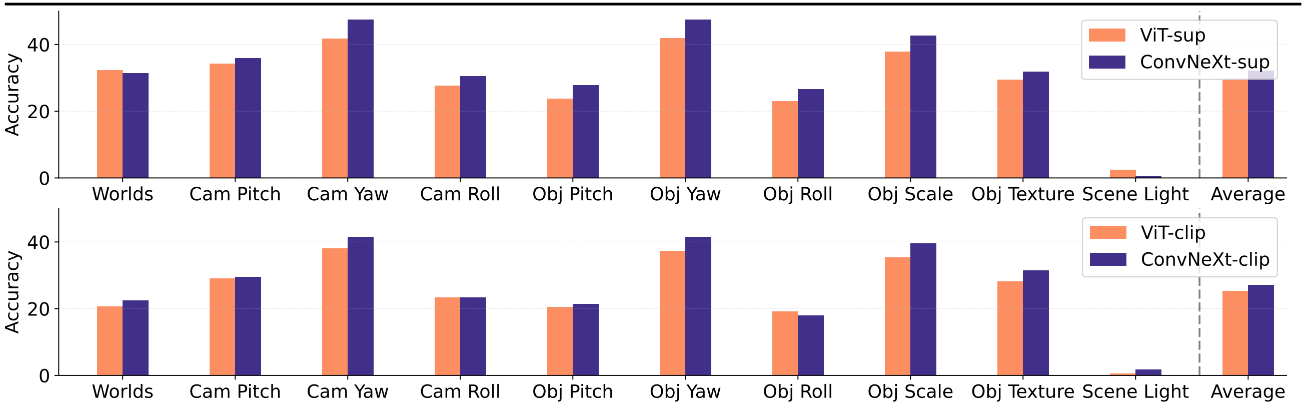


Figure 7. **Results on synthetic data from PUG-ImageNet.** ConvNeXt is superior on almost every factor for both supervised and CLIP.

evaluate how well this model performs when applied to a vastly different domain, such as medical imaging.

To assess the transferability, we adopted a VTAB benchmark (Zhai et al., 2019). It comprises 19 diverse datasets grouped into three subcategories: natural, specialized, and structured. We conduct a linear probing evaluation on frozen features, following the protocol from (Ilharco et al., 2021). The results are shown in Fig. 6 (bottom) and Table 2.

Model	Natural	Specialized	Structured	Overall
ViT-sup	84.2	84.2	45.4	67.8
ConvNeXt-sup	87.1	85.0	50.0	71.0
ViT-clip	87.6	87.8	50.9	72.2
ConvNeXt-clip	87.8	86.9	51.2	72.2

Table 2. **Transferability results on VTAB in subgroups.** CLIP models are better on each of the dataset subgroups. For supervised models, ConvNeXt outperforms ViT by a large margin.

Supervised ConvNeXt has great transferability, almost matching the performance of CLIP models. We find that ConvNeXt strongly outperforms ViT for supervised. Interestingly the performance of supervised ConvNeXt is not very far from CLIP models, both of which have the same average accuracy. For CLIP, ViT and ConvNeXt demonstrate similar average accuracy, with many datasets showing a performance gap of less than 1%. CLIP models generally show better transferability on all three subgroups of VTAB (Table 2), which is different from the robustness experiments. The superiority of CLIP can be attributed to the larger and more diverse volume of pretraining data (Ramanujan et al., 2023).

3.6. Synthetic Data

While two previous sections focused on robustness and transferability, they did not cover the new and promising area of training models with synthetic data (Tian et al., 2023). Unlike human-annotated data, synthetic datasets allow precise control over the content and quality of data. PUG-ImageNet (Bordes et al., 2023) is a synthetic dataset of photorealistic images of ImageNet classes that provides labels for a set of factors. The images are generated using a software that allows systematically varying factors like pose, size, texture, and others for each object. In our experiments, we provide top-1 accuracy results for ten different factors in PUG-ImageNet and their average in Fig. 7.

ConvNeXt is better than ViT on synthetic data. Intriguingly, ConvNeXt outperforms ViT on PUG-ImageNet for nearly all factors. This suggests: ConvNeXt is better than ViT on synthetic data. CLIP models have lower accuracy compared to supervised, which is likely related to their inferior performance on the original ImageNet.

3.7. Transformation Invariance

In real-world scenarios, data often undergo transformations that preserve its semantic meaning or class. We aim to ensure that the model's representations are invariant to these transformations. Achieving various types of invariance is desirable because it enables the network to generalize well across different but semantically similar inputs, thereby enhancing its robustness and predictive power. In previous literature (Azulay & Weiss, 2018; Zhang, 2019), it has been shown that the performance of neural networks can be highly unstable even under simple input data transformations, such as shifting an input image by a few pixels.

We conduct experiments to assess three types of invariance: scale, shift, and resolution. We analyze the model's accuracy trends on the ImageNet-1K validation set as a function of varying scale / shift magnitude and image resolution. In scale invariance analysis, the image is first resized according to a given scale factor, and then a central crop is taken. In shift experiments, we adjust the crop location in the original image space and then take a crop, shifting along the longer side of the image. In resolution experiments with ViT model, we interpolate positional embeddings to match the new applied resolution.

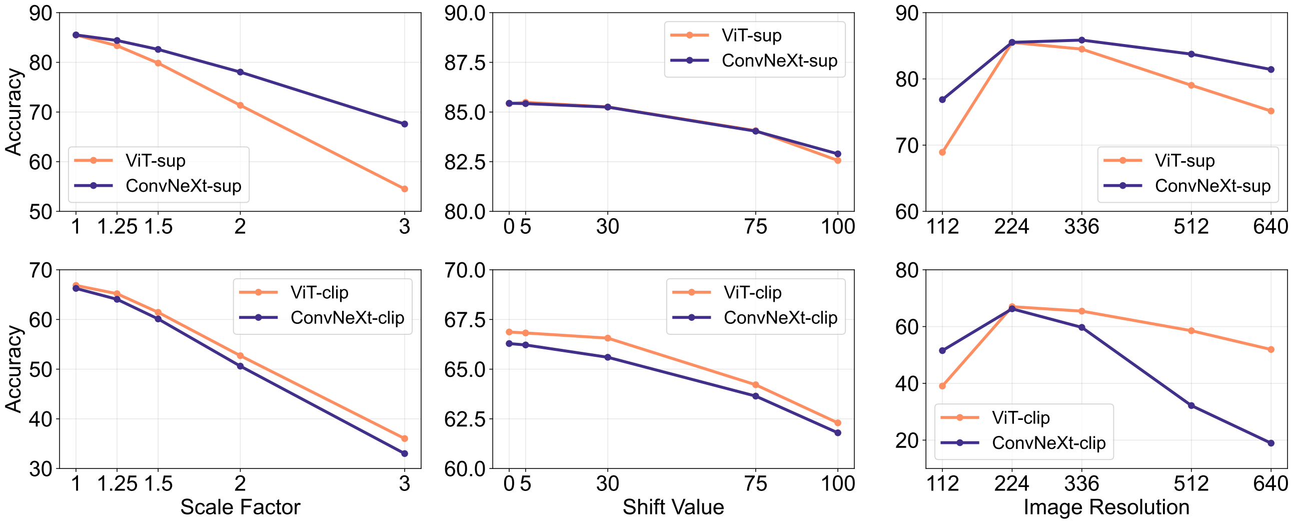


Figure 8. **Scale, shift, and resolution invariance experiments.** ConvNeXt is better than ViT under supervised training on all transformation types. All models are robust to shift transformation but experience degradation when the image scale is altered.

Supervised ConvNeXt is the most invariant model to the data transformations. We display our results in Fig. 8, observing a consistent trend of ConvNeXt outperforming ViT under supervised training. Interestingly, supervised ConvNeXt has better performance on 336 pixel resolution than on the original resolution of 224 pixels. Overall, all models are robust to shifting and less robust to scaling and resolution transforms. For practical use cases requiring high transform invariance, our results indicate that supervised ConvNeXt will be the best choice among analyzed models.

4. Related Work

Architecture analysis. Several works compared ViTs and ConvNeXt from the perspective of internal representations (Raghu et al., 2021), synthetic data (Ruiz et al., 2022), transferability (Zhou et al., 2021), and robustness (Wang et al., 2022; Bai et al., 2021; Pinto et al., 2022; Djolonga et al., 2021). Other studies included analysis of Transformer properties (Naseer et al., 2021) and impact of neural network width and depth on learned representations (Nguyen et al., 2020). ViTs and ConvNets were also evaluated on ImageNet, showing that Transformers are more aligned with human error patterns (Tuli et al., 2021). A large variety of backbones, trained with various methods, were benchmarked in (Goldblum et al., 2024) across a diverse set of computer vision tasks, including classification, detection, and retrieval. In contrast to studies that analyze a single property, our work extensively compares models across many, maintaining a fair comparison by evaluating models with similar ImageNet accuracies.

Training objective analysis. A comprehensive analysis was conducted in (Walmer et al., 2023), comparing ViTs trained with supervised, self-supervised, and CLIP objectives. Analysis of the representations of models trained with supervised and self-supervised objectives was presented in (Grigg et al., 2021; Gwilliam & Shrivastava, 2022). Two works (Park et al., 2023; Shekhar et al., 2023) focused on investigating the effect of training objective in self-supervised learning. Unlike studies emphasizing self-supervised models, our work compares supervised and CLIP models.

Limitations of ImageNet. Recent research (Beyer et al., 2020; Recht et al., 2019; Tsipras et al., 2020; Yun et al., 2021) highlighted issues with the reliability and quality of ImageNet labels. Two studies (Kornblith et al., 2019; Miller et al., 2021) showed a strong relationship between performance on ImageNet and other datasets, although this can depend on the model's architecture and training methods. Other studies (Richards et al., 2023; Fang et al., 2023b) showed that high ImageNet accuracy does not ensure good performance on diverse datasets. Current robustification training techniques were found to overfit (Yamada & Otani, 2022) to ImageNet evaluations. In addition, ImageNet suffers from dichotomous data difficulty (Meding et al., 2021), obscuring differences between models. Our analysis does not directly address data-related problems of ImageNet but instead studies alternative properties.

5. Conclusion

Our study examined ConvNets and Transformers with supervised and CLIP training from multiple perspectives beyond the standard ImageNet accuracy. We found that models with similar ImageNet accuracies have vastly different properties. This suggests that model selection should depend on the target use cases, as standard metrics may overlook key nuances. In addition, it is crucial to develop new benchmarks with data distributions that closely mirror real-world scenarios. This will help both in training models for better real-world performance and in more accurately evaluating their effectiveness in such environments.

ConvNet vs Transformer. (1) Supervised ConvNeXt is superior to supervised ViT: it is more invariant to data transformations, and demonstrates better transferability, robustness and calibration. (2) ConvNeXt outperforms ViT on synthetic data. (3) ViT has a higher shape bias.

Supervised vs CLIP. (1) Supervised ConvNeXt competes well with CLIP in transferability, showing potential of supervised models. (2) Supervised models are better at robustness benchmarks, likely because these are ImageNet variants. (3) CLIP models have a higher shape bias and make less classification errors relative to their ImageNet accuracy.

As a result of our analysis, we suggest using supervised ConvNeXt when the target task distribution is not very different from ImageNet as this model provides competitive performance among many benchmarks. In case of a serious domain shift, we recommend using CLIP models.

Acknowledgments. We would like to thank Sheng Zhang, Kudaibergen Abutalip, Guangyi Chen, Nurdaulet Mukhituly, Boyang Sun, Zhiqiu Xu, and Nurbek Tastan for their valuable feedback and suggestions. Use of all datasets and all experiments conducted by MBZUAI.

Impact Statement

This paper presents work whose goal is to advance the fields of Machine Learning, Deep Learning and Computer Vision. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Azulay, A. and Weiss, Y. Why do deep convolutional networks generalize so poorly to small image transformations? *arXiv preprint arXiv:1805.12177*, 2018.

Ba, J. L., Kiros, J. R., and Hinton, G. E. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Bai, Y., Mei, J., Yuille, A. L., and Xie, C. Are transformers more robust than cnns? In *NeurIPS*, 2021.

Beyer, L., Hénaff, O. J., Kolesnikov, A., Zhai, X., and Oord, A. v. d. Are we done with imagenet? *arXiv preprint arXiv:2006.07159*, 2020.

Bordes, F., Shekhar, S., Ibrahim, M., Bouchacourt, D., Vincent, P., and Morcos, A. S. Pug: Photorealistic and semantically controllable synthetic data for representation learning. *arXiv preprint arXiv:2308.03977*, 2023.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *ICML*, 2020.

Dehghani, M., Djolonga, J., Mustafa, B., Padlewski, P., Heek, J., Gilmer, J., Steiner, A. P., Caron, M., Geirhos, R., Alabdulmohsin, I., et al. Scaling vision transformers to 22 billion parameters. In *ICML*, 2023.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

Djolonga, J., Yung, J., Tschannen, M., Romijnders, R., Beyer, L., Kolesnikov, A., Puigcerver, J., Minderer, M., D'Amour, A., Moldovan, D., et al. On robustness and transferability of convolutional neural networks. In *CVPR*, 2021.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Fang, A., Jose, A. M., Jain, A., Schmidt, L., Toshev, A., and Shankar, V. Data filtering networks, 2023a.

Fang, A., Kornblith, S., and Schmidt, L. Does progress on imagenet transfer to real-world datasets? *arXiv preprint arXiv:2301.04644*, 2023b.

Gadre, S. Y., Ilharco, G., Fang, A., Hayase, J., Smyrnis, G., Nguyen, T., Marten, R., Wortsman, M., Ghosh, D., Zhang, J., et al. Datacomp: In search of the next generation of multimodal datasets. *arXiv preprint arXiv:2304.14108*, 2023.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., and Brendel, W. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2020.

Goldblum, M., Souri, H., Ni, R., Shu, M., Prabhu, V., Somepalli, G., Chattopadhyay, P., Ibrahim, M., Bardes, A., Hoffman, J., et al. Battle of the backbones: A large-scale comparison of pretrained models across computer vision tasks. *NeurIPS*, 2024.

Grigg, T. G., Busbridge, D., Ramapuram, J., and Webb, R. Do self-supervised and supervised methods learn similar visual representations? *arXiv preprint arXiv:2110.00528*, 2021.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *ICML*, 2017.

Gwilliam, M. and Shrivastava, A. Beyond supervised vs. unsupervised: Representative benchmarking and analysis of image representation learning. In *CVPR*, 2022.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *CVPR*, 2016.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. Masked autoencoders are scalable vision learners. In *CVPR*, 2022.

Hendrycks, D. and Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *CVPR*, 2021a.

Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., and Song, D. Natural adversarial examples. In *CVPR*, 2021b.

Idrissi, B. Y., Bouchacourt, D., Balestriero, R., Evtimov, I., Hazirbas, C., Ballas, N., Vincent, P., Drozdzal, M., Lopez-Paz, D., and Ibrahim, M. Imagenet-x: Understanding model mistakes with factor of variation annotations. *arXiv preprint arXiv:2211.01866*, 2022.

Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., and Schmidt, L. Openclip, July 2021.

Kolesnikov, A., Beyer, L., Zhai, X., Puigcerver, J., Yung, J., Gelly, S., and Houlsby, N. Big transfer (bit): General visual representation learning. In *ECCV*, 2020.

Kornblith, S., Shlens, J., and Le, Q. V. Do better imagenet models transfer better? In *CVPR*, 2019.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. 2012.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., and Xie, S. A convnet for the 2020s. In *CVPR*, 2022.

Luo, H., Ji, L., Zhong, M., Chen, Y., Lei, W., Duan, N., and Li, T. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 2022.

Meding, K., Buschoff, L. M. S., Geirhos, R., and Wichmann, F. A. Imagenet suffers from dichotomous data difficulty. In *NeurIPS 2021 Workshop on ImageNet: Past, Present, and Future*, 2021.

Miller, J. P., Taori, R., Raghunathan, A., Sagawa, S., Koh, P. W., Shankar, V., Liang, P., Carmon, Y., and Schmidt, L. Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In *ICML*, 2021.

Minderer, M., Djolonga, J., Romijnders, R., Hubis, F., Zhai, X., Houlsby, N., Tran, D., and Lucic, M. Revisiting the calibration of modern neural networks. *NeurIPS*, 2021.

Naseer, M. M., Ranasinghe, K., Khan, S. H., Hayat, M., Shahbaz Khan, F., and Yang, M.-H. Intriguing properties of vision transformers. In *NeurIPS*, 2021.

Nguyen, T., Raghu, M., and Kornblith, S. Do wide and deep networks learn the same things? uncovering how neural network representations vary with width and depth. *arXiv preprint arXiv:2010.15327*, 2020.

Park, N., Kim, W., Heo, B., Kim, T., and Yun, S. What do self-supervised vision transformers learn? *arXiv preprint arXiv:2305.00729*, 2023.

Pinto, F., Torr, P. H., and K. Dokania, P. An impartial take to the cnn vs transformer robustness contest. In *ECCV*, 2022.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C., and Dosovitskiy, A. Do vision transformers see like convolutional neural networks? *NeurIPS*, 2021.

Ramanujan, V., Nguyen, T., Oh, S., Schmidt, L., and Farhadi, A. On the connection between pre-training data diversity and fine-tuning robustness. *arXiv preprint arXiv:2307.12532*, 2023.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Do imagenet classifiers generalize to imagenet? In *ICML*, 2019.

Richards, M., Kirichenko, P., Bouchacourt, D., and Ibrahim, M. Does progress on object recognition benchmarks improve real-world generalization? *arXiv preprint arXiv:2307.13136*, 2023.

Ruiz, N., Bargal, S., Xie, C., Saenko, K., and Sclaroff, S. Finding differences between transformers and convnets using counterfactual simulation testing. *NeurIPS*, 2022.

Shekhar, S., Bordes, F., Vincent, P., and Morcos, A. Objectives matter: Understanding the impact of self-supervised objectives on vision transformer representations. *arXiv preprint arXiv:2304.13089*, 2023.

Taesiri, M. R., Nguyen, G., Habchi, S., Bezemer, C.-P., and Nguyen, A. Imagenet-hard: The hardest images remaining from a study of the power of zoom and spatial biases in image classification.

Tian, Y., Fan, L., Isola, P., Chang, H., and Krishnan, D. Stablerep: Synthetic images from text-to-image models make strong visual representation learners. *arXiv preprint arXiv:2306.00984*, 2023.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jégou, H. Training data-efficient image transformers & distillation through attention. In *ICML*, 2021.

Touvron, H., Cord, M., and Jégou, H. Deit iii: Revenge of the vit. In *ECCV*, 2022.

Tsipras, D., Santurkar, S., Engstrom, L., Ilyas, A., and Madry, A. From ImageNet to image classification: Contextualizing progress on benchmarks. In *ICML*, 2020.

Tuli, S., Dasgupta, I., Grant, E., and Griffiths, T. L. Are convolutional neural networks or transformers more like human vision? *arXiv preprint arXiv:2105.07197*, 2021.

Vinker, Y., Alaluf, Y., Cohen-Or, D., and Shamir, A. Clipascene: Scene sketching with different types and levels of abstraction. In *ICCV*, 2023.

Walmer, M., Suri, S., Gupta, K., and Shrivastava, A. Teaching matters: Investigating the role of supervision in vision transformers. In *CVPR*, 2023.

Wang, H., Ge, S., Lipton, Z., and Xing, E. P. Learning robust global representations by penalizing local predictive power. *NeurIPS*, 2019.

Wang, Z., Bai, Y., Zhou, Y., and Xie, C. Can cnns be more robust than transformers? *arXiv preprint arXiv:2206.03452*, 2022.

Woo, S., Debnath, S., Hu, R., Chen, X., Liu, Z., Kweon, I. S., and Xie, S. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *CVPR*, 2023.

Wortsman, M., Ilharco, G., Kim, J. W., Li, M., Kornblith, S., Roelofs, R., Lopes, R. G., Hajishirzi, H., Farhadi, A., Namkoong, H., et al. Robust fine-tuning of zero-shot models. In *CVPR*, 2022.

Yamada, Y. and Otani, M. Does robustness on imagenet transfer to downstream tasks? In *CVPR*, 2022.

Yun, S., Oh, S. J., Heo, B., Han, D., Choe, J., and Chun, S. Re-labeling imagenet: from single to multi-labels, from global to localized labels. In *CVPR*, 2021.

Zhai, X., Puigcerver, J., Kolesnikov, A., Ruyssen, P., Riquelme, C., Lucic, M., Djolonga, J., Pinto, A. S., Neumann, M., Dosovitskiy, A., et al. A large-scale study of representation learning with the visual task adaptation benchmark. *arXiv preprint arXiv:1910.04867*, 2019.

Zhang, R. Making convolutional networks shift-invariant again. In *ICML*, 2019.

Zhou, H.-Y., Lu, C., Yang, S., and Yu, Y. Convnets vs. transformers: Whose visual representations are more transferable? In *CVPR*, 2021.

Appendix

A. Results for Models of Different Sizes

Besides having the base sized models in the main part of our analysis, we additionally include the following models to see the effect of the model size on the performance:

- Supervised ConvNeXt: Tiny, Small, Large, Huge (XLarge).
- Supervised ViT (DeiT 3): Small, Large, Huge (XLarge). Note that no Tiny version is provided for DeiT 3 by the original authors (Touvron et al., 2022).
- CLIP ConvNeXt: Large, Huge (XLarge).
- CLIP ViT: Large, Huge.

All new additional CLIP models are pretrained on the LAION-2B dataset, while the CLIP models from the main part of the manuscript (Table 1) are pretrained on LAION-400M. All CLIP models are taken from OpenCLIP, and supervised models are taken from the respective original works (Liu et al., 2022; Touvron et al., 2022).

Architecture	Pretraining	Finetuning	Paradigm	FLOPs	#Params	INet-1K val%
ConvNeXt-Tiny	ImageNet-21K	ImageNet-1K	Supervised	4.5G	29M	82.9
ConvNeXt-Small	ImageNet-21K	ImageNet-1K	Supervised	8.7G	50M	84.6
ConvNeXt-Large	ImageNet-21K	ImageNet-1K	Supervised	34.4G	198M	86.6
ConvNeXt-Huge	ImageNet-21K	ImageNet-1K	Supervised	60.9G	350M	87.0
ViT-S/16	ImageNet-21K	ImageNet-1K	Supervised	4.6G	22M	83.1
ViT-L/16	ImageNet-21K	ImageNet-1K	Supervised	61.6G	304M	87.0
ViT-H/14	ImageNet-21K	ImageNet-1K	Supervised	167.4G	632M	87.2
ConvNeXt-Large	LAION-2B	—	CLIP	34.4G	198M	75.2
ConvNeXt-Huge	LAION-2B	—	CLIP	60.9G	350M	78.6
ViT-L/14	LAION-2B	—	CLIP	61.6G	304M	74.8
ViT-H/14	LAION-2B	—	CLIP	167.4G	632M	77.3

Table 3. **Additional models configurations.** We perform partial analysis for the additional models of different sizes to see the effect of the model size on the performance.

Additional models are evaluated on PUG-ImageNet (Table 4), ImageNet-X (Table 5), calibration (Table 6 and 7), transformation invariance (Table 8), and shape / texture bias (Table 9). Based on the results from new models we make the following observations:

- On PUG-ImageNet (Table 4) ConvNeXt is better than ViT in 6 out of 7 comparison, suggesting its clear advantage over ViT on synthetic data.
- ImageNet-X performance is largely determined by the training method (Table 5). CLIP models are clearly better than supervised.
- On calibration (Table 6 and 7) ConvNeXt has lower ECE value compared to its ViT counterpart in most cases. This solidifies our initial conclusion in the main part that ConvNeXt is better calibrated than ViT.
- Shape bias greatly improves with scale (Table 9). Interestingly, the ConvNeXt is better than ViT on large-scale CLIP models (Large and Huge). For Huge CLIP models ConvNeXt has an advantage of almost 10% over ViT. This suggests that training method and model size has a noticeable influence on the shape bias of the model. Moreover, this result also highlights that ConvNeXt has the potential to exceed a ViT on this benchmark.
- Even small supervised models are robust to shift transformation (Table 8 middle part). For example, the smallest supervised model ConvNeXt-Tiny has a tiny degradation of $< 3\%$ which is comparable to the huge CLIP models.
- Huge supervised models are quite reliable to the resolution change (Table 8 right part).
- In general, large models provide a decent improvement over the base models. However, huge models provide only marginal improvement over the large models.

Model	Tiny	Small	Base	Large	Huge
ConvNeXt-sup	27.0	29.4	32.2	34.6	35.6
ViT-sup	—	24.7	29.4	34.3	36.9
ConvNeXt-clip	—	—	27.0	35.4	43.2
ViT-clip	—	—	25.2	32.9	38.8

Table 4. **PUG-ImageNet results.** ConvNeXt is better in 6 out of 7 comparisons.

Model	Tiny	Small	Base	Large	Huge
ConvNeXt-sup	1.48	1.50	1.43	1.47	1.50
ViT-sup	—	1.46	1.45	1.49	1.50
ConvNeXt-clip	—	—	1.23	1.26	1.32
ViT-clip	—	—	1.19	1.33	1.31

Table 5. **ImageNet-X results.** CLIP models have an advantage.

Model	Tiny	Small	Base	Large	Huge
ConvNeXt-sup	0.021	0.022	0.025	0.029	0.031
ViT-sup	—	0.034	0.043	0.047	0.041
ConvNeXt-clip	—	—	0.085	0.017	0.019
ViT-clip	—	—	0.073	0.030	0.031

Table 6. **Calibration on ImageNet-1K**. ConvNeXt has lower ECE in most cases.

Model	Tiny	Small	Base	Large	Huge
ConvNeXt-sup	0.073	0.035	0.028	0.022	0.023
ViT-sup	—	0.056	0.041	0.037	0.037
ConvNeXt-clip	—	—	0.141	0.080	0.090
ViT-clip	—	—	0.133	0.094	0.098

Table 7. **Calibration on ImageNet-R.** ConvNeXt is better calibrated than ViT.

Model	Scale Invariance					Shift Invariance					Resolution Invariance				
	1x	1.25x	1.5x	2x	3x	0px	5px	30px	75px	100px	112px	224px	336px	512px	640px
ConvNeXt-sup Tiny	82.9	81.8	79.5	74.6	63.2	82.9	82.9	82.6	81.4	80.1	67.2	82.9	77.2	42.8	21.1
ConvNeXt-sup Small	84.6	83.3	81.4	76.8	66.4	84.6	84.5	84.5	83.3	82.1	74.8	84.6	84.7	82.5	80.6
ConvNeXt-sup Base	85.5	84.5	82.7	78.3	68.5	85.5	85.7	85.6	84.5	83.2	78.1	85.5	85.8	83.7	81.5
ConvNeXt-sup Large	86.6	85.4	84.9	79.9	70.7	86.6	86.6	86.5	85.4	84.5	80.2	86.6	86.2	84.6	82.9
ConvNeXt-sup Huge	87.0	85.8	84.3	80.4	71.5	87.0	86.8	86.7	85.7	84.7	81.0	87.0	86.7	84.8	82.9
ViT-sup Small	82.7	80.5	76.4	65.7	48.2	82.7	82.8	82.5	81.1	79.2	65.5	82.7	81.7	75.4	70.9
ViT-sup Base	85.5	83.3	79.7	70.6	54.8	85.5	85.5	85.2	84.1	82.6	69.5	85.5	84.2	78.1	74.5
ViT-sup Large	86.8	85.0	81.9	74.2	60.1	86.8	86.8	86.6	85.5	84.3	76.7	86.8	85.9	81.7	79.1
ViT-sup Huge	86.9	85.2	81.8	73.9	59.6	86.9	86.8	86.7	85.7	84.5	73.1	86.9	85.9	81.7	79.4
ConvNeXt-clip Base	66.3	64.0	60.0	50.4	33.3	66.3	66.2	65.7	63.6	61.8	51.5	66.3	59.7	32.2	18.9
ConvNeXt-clip Large	75.2	73.5	71.3	64.9	50.6	75.2	75.1	74.8	73.5	72.1	62.1	75.2	76.1	70.3	64.3
ConvNeXt-clip Huge	78.6	77.3	75.2	69.6	56.3	78.6	78.6	78.2	77.1	75.8	68.0	78.6	78.8	71.0	61.9
ViT-clip Base	67.0	65.0	61.3	52.5	36.1	67.0	66.9	66.6	64.2	62.2	39.1	67.0	65.4	58.5	51.9
ViT-clip Large	74.8	72.1	68.1	58.7	41.4	74.8	74.7	74.3	72.0	69.6	48.0	74.8	74.4	69.5	64.9
ViT-clip Huge	77.3	74.9	71.4	62.7	45.5	77.3	77.4	76.9	74.7	72.7	51.1	77.3	76.9	72.7	68.4

Table 8. **Scale, shift, and resolution invariance experiments.**

Model	Tiny	Small	Base	Large	Huge
ConvNeXt-sup	28.1	31.6	33.3	39.8	41.3
ViT-sup	—	35.0	39.8	51.3	57.1
ConvNeXt-clip	—	—	45.9	56.9	67.0
ViT-clip	—	—	46.4	53.9	58.5

Table 9. **Shape-texture bias results.** Shape bias noticeably improves with model size.