Models trained on datasets such as ImageNet, often struggle to transfer their performance to real-world applications where conditions and scenarios are significantly more diverse.(occlusions, diffrnt light conditions..)

CLIP models, excel in generalization and transferability and offer intriguing representational properties that connect vision and language. CLIP ’ s vision encoder exhibits significantly better robustness and transferability.
CLIP models make fewer classification errors relative to their ImageNet performance.
CLIP models have smaller texture bias than supervised counter parts(CLIP Vit CLIP ConvNext almost similar and ViTs exhibit stronger shape bias than ConvNeXts for both supervised and CLIP models).
However, supervised models are better calibrated(predictions confidence reflect accuracy) and Supervised ConvNeXt is better calibrated than supervised ViT.
Supervised models are generally superior on ImageNet robustness benchmarks(Imanget distribution shifts - ImageNet-V2, ImageNet-A, ImageNet-C, ImageNet-R, ImageNet-Sketch, ImageNet-Real and ImageNet-Hard) 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 the supervised models were finetuned. ViT and ConvNeXt, on average, exhibit similar performance across both supervised and CLIP in robustness.
Clip has better transferibility.Supervised ConvNeXt has great transferability almost matching the performance of CLIP models. Supervised ConvNeXt strongly outperforms ViT and performance of supervised models is not very far from CLIP models, both of which have the same average accuracy. For CLIP models, ViT and ConvNeXt demonstrate similar average performance and show better transferability on all three subgroups (Table 2) of VTAB, which is different from the robustness experiments. Their superiority can be attributed to the larger and more diverse pretraining data
ConvNeXt is better than ViT when trained on synthetic data and for CLIP models, the gap between ConvNeXt and ViT is slightly smaller than that for supervised models, and they generally have lower accuracy compared to supervised models. This is likely related to their inferior accuracy on the original ImageNet.
Supervised ConvNeXt excels in transformation invariance. ConvNeXt outperforming ViT under supervised training. This trend is reversed for CLIP models, likely because ConvNeXt-clip was undertrained. Overall, models are reliable to shift transformation and less robust to scale and resolution transforms. For practical applications requiring high robustness to scale, shift, and resolution transforms, our results indicate that ConvNeXt under supervised training could be the best choice.
ConvNeXt has an advantage on synthetic data but is more texture biased than ViT(For both supervised and clip based trainings, more shape bias is what humans usually perform recognitiion in the real world). Supervised ConvNeXt excels on many benchmarks and has transferability performance close to CLIP models, suggesting that architecture is still a crucial factor.

ConvNet vs Transformer, Supervised vs CLIP: Beyond ImageNet Accuracy

Selecting the model/model arch for our application should not just only be based on one single metric i.e their Imagenet classification accuracy performance, as they all differ in different characterstics/ performance nuances/aspects like model mistakes (incorrect predictions errors), shape and texture bias ,output calibration performance, robustness, transferability, generalization capabilities, feature transformation invariance(invariances of the learned representations), training models on synthetic data

Diversity in model characteristics, not captured by traditional metrics, highlights the need for more detailed evaluation metrics for accurate, context-specific model selection and the development of new benchmarks non-related to ImageNet

Kirill

Zhuang Liu[2]

ch

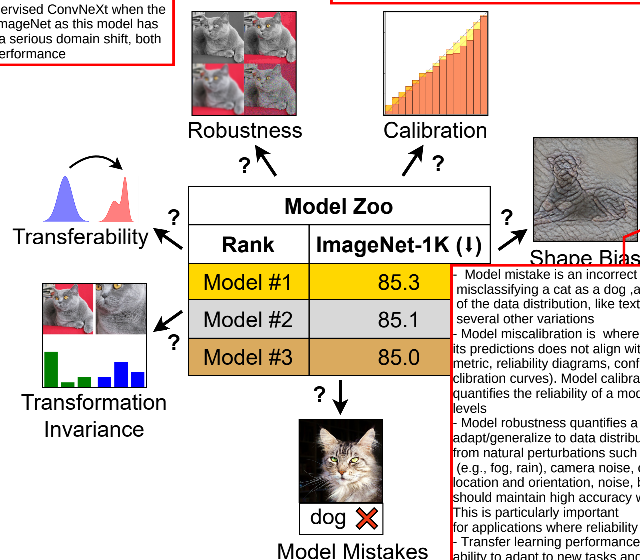
For supervised models, we found the superior performance of ConvNeXt over ViT on many benchmarks. It is better calibrated, more invariant to data transformation, and has shown superior transferability and robustness. More over, both CLIP and supervised ConvNeXt are better than ViT on synthetic data.
While CLIP models have lower ImageNet accuracy, they exhibit higher shape bias and better transferability, making them preferable in scenarios with significant domain shifts. Finally, all models make largely the same types of mistakes and struggle with texture.
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 has the strongest performance among all. In case of a serious domain shift, both of the CLIP models should provide competitive performance

Comparing Transformers Vit and CNN Convnext in 2 training paradigms and all these models have similar Imagenet Accuracy(for clip based mdls, its zero-shot accuracy) and compute requirements. One is both the models Vit and ConvNext in supervised setting and the other is both the models in CLIP image text training setting. Self supervised training paradigm(MAE) is not tested here as they show behaviors similar to supervised models and this could be due to their final ImageNet-1K supervised finetuning

For supervised models, we use a pretrained DeiT3-Base/16 for ViT, which shares the same architecture as ViT-Base/16 with an improved training recipe, and ConvNeXt-Base.
For CLIP models, we use vision encoders of ViT-Base/16 and ConvNeXt-Base from OpenCLIP instead of openAI

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. Our code is available at https://github.com/kirill-vish/Beyond-INet.



- Model mistake is an incorrect label assignment, such as misclassifying a cat as a dog ,according to certain fcators of the data distribution, like texture,occlusions, shape and several other variations
- Model miscalibration is where a model's confidence in its predictions does not align with actual accuracy(ECE metric, reliability diagrams, confidence histograms, clibration curves). Model calibration is a metric that quantifies the reliability of a model's predicted confidence levels
- Model robustness quantifies a model's capability to adapt/generalize to data distribution shifts that can arise from natural perturbations such as atmospheric conditions (e.g., fog, rain), camera noise, or variations in object location and orientation, noise, blur). A robust model should maintain high accuracy with these perturbations. This is particularly important for applications where reliability is a primary concern
- Transfer learning performance of a model indicates its ability to adapt to new tasks and datasets beyond its original training domain.(consider a model that has been originally trained on ImageNet, which primarily consists of natural images. A test of its transferability would be to evaluate how well this model performs when applied to a vastly different domain,such as medical imaging)
- Models representations are invariant to shift, scale and resolution transformations.

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

1. Introduction

The computer vision model landscape has become increasingly complex. From early ConvNets [30] to advances in Vision Transformers [11], the variety of models available has expanded significantly. Similarly, training paradigms have evolved from supervised training on ImageNet [9] to self-supervised learning [7, 19] and image-text pair training like CLIP [40]. While signaling progress, this explosion of choices poses a significant challenge for practitioners: selecting a model that suits their purposes.

SIMCLR, MOCO

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 [29]. 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 [5].

Are we done with imagenet?

A particularly noteworthy example is CLIP. Despite having a similar ImageNet accuracy as a ResNet [18], CLIP's vision encoder exhibits significantly better robustness and transferability. This has sparked research that explores and builds upon the unique strengths of CLIP [32, 43, 54, 59], that could not be exposed by the ImageNet metric alone. This illustrates that analyzing alternative properties could help discover useful models.

In addition to curious scientific inquiries, 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 conditions, or occlusions. For instance, models trained on datasets such as ImageNet, often struggle [60] to transfer their performance to real-world applications where

An empirical study of clip for end to end video clip retrieval and captioning, Hierarchical text-conditional image generation with clip latents, Clipascene: Scene sketching with different types and levels of abstraction, Robust fine-tuning of zero-shot models.

Does robustness on imagenet transfer to downstream tasks?

Masked autoencoders are scalable vision learners
Co-designing and scaling convnets with masked autoencoders

Model	Architecture	Pretraining	Finetuning	Paradigm	FLOPs	#Param	INet-1K val%
ViT-sup	ViT-B/16	ImageNet-21K	ImageNet-1K	supervised	17.5G	87M	83.7
ConvNeXt-sup	ConvNeXt-B	ImageNet-21K	ImageNet-1K	supervised	15.4G	89M	83.7
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.

Scaling vision transformers to 22 billion parameters

conditions and scenarios are significantly 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 computer vision: ConvNeXt [31], as a representative ConvNet, and Vision Transformer (ViT) [11], each with supervised and CLIP training paradigms. 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 performance, and many others. Notably, we focus on properties that the model exhibits without further training or finetuning, so that they can guide practitioners who wish to use 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 generally 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 has transferability performance close to CLIP models, suggesting that architecture is still a crucial factor. 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 development of new benchmarks non-related to ImageNet.

Are transformers more robust than cnns?
Revisiting the calibration of modern neural networks
Intriguing properties of vision transformers.
Convnets vs. transformers: Whose visual representations are more transferable?

2. Models

For analyzing ConvNets and Transformers, many previous works [4, 35, 36, 64] compare ResNet and ViT, making the comparison 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 [3] that were not incorporated in ResNet when it was invented years ago. For a more balanced evaluation, we compare ViT with ConvNeXt [31], a modern representative of ConvNet that matches Transformers' performance and shares many of their designs.

As for the training paradigms, we compare supervised and CLIP. Supervised models continue to show state-of-the-art performance in computer vision [8]. CLIP models, on the other hand, excel in generalization and transferability and offer intriguing representational properties that connect vision and language. Self-supervised models [20, 58] are not included in the results as they show 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. They also have similar sizes/compute 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 [51] for ViT, which shares the same architecture as ViT-Base/16 with an improved training recipe, and ConvNeXt-Base [31]. For CLIP models, we use vision encoders of ViT-Base/16 and ConvNeXt-Base from OpenCLIP [26]. Note that these models have a slightly different performance from the original OpenAI models [40]. All model checkpoints can be found in our GitHub repo. A detailed model comparison is given in Table 1.

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 computational resources, who often depend 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. Following this, we move on to give a detailed analysis of different properties individually.

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.

Imagenet-X dataset offers detailed human annotations for 16 factors of variation, such as pose, style, and others
For model prediction errors w,r.to factors, CLIP models make fewer mistakes relative to their ImageNet accuracy than supervised.,
CLIP models are much more robust towards shape, subcategory, texture, object blocking, and darker factors compared to supervised models as clip models trained on diverse data.
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, shape,subcategory, texture and occlusions like object blocking, person blocking, multiple objects
Texture is the most challenging factor for all models suggests that models of the current generation largely suffer because of texture bias

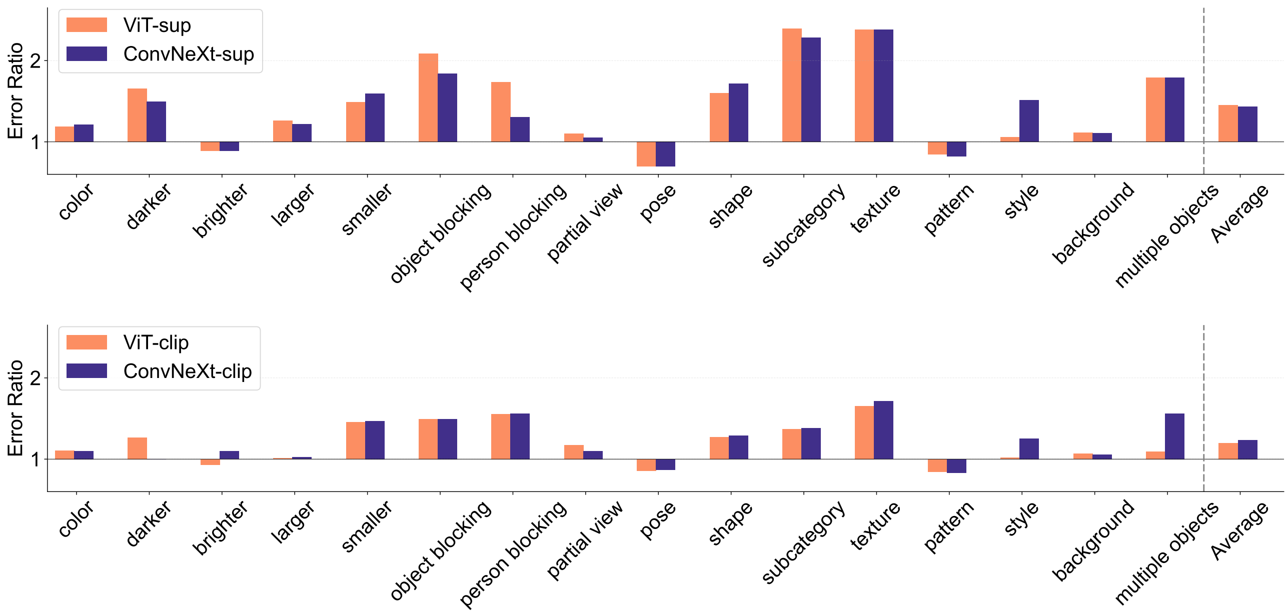


Figure 2. **Model mistakes on ImageNet-X.** Lower is better. ConvNeXt and ViT models perform similarly within each training category. CLIP models achieve lower error ratios.

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, offering advantages over a black-box approach.

The ImageNet-X dataset [25] offers detailed human annotations for 16 factors of variation, such as pose, style, and others. This allows a targeted 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 for selected models 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's important to note that the error ratio is relative to overall ImageNet accuracy, where a significant 16% 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 difference between CLIP and supervised models is likely the more diverse training data used for CLIP.

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 instance, in occlusion, the model often misclassifies due to focusing on the visible, obscuring object.

Texture is the most challenging factor for all models. Interestingly, all models in our analysis make 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. A more detailed analysis of shape / texture bias is provided in the next Section 3.2.

3.2. Shape / Texture Bias

In contrast to humans, who generally use high-level visual cues for recognition, neural networks often rely on more

The study of shape-texture bias serves to highlight the phenomenon of neural networks often rely on more brittle shortcut features 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 accounts for the proportion favoring the class represented by the texture. These metrics reveal whether the classifier favors shape or texture when they conflict. ConvNets have a strong bias towards texture, as opposed to shape, which differs from human behavior. For both supervised and clip based trainings, ViT is less biased towards the texture than ConvNet. Notably, scaling large Transformer models has led to shape biases comparable to human level.
CLIP models have smaller texture bias than supervised.

Dashed lines represent average shape bias aggregated over all the categories. Individual markers on horizontal lines depict shape bias for the particular class represented with a 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

we can observe that ViTs exhibit stronger shape bias than ConvNeXts for both supervised and CLIP models, possibly because ConvNeXt is more inclined to learn local features related to textures due to the local nature of convolution operations. 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

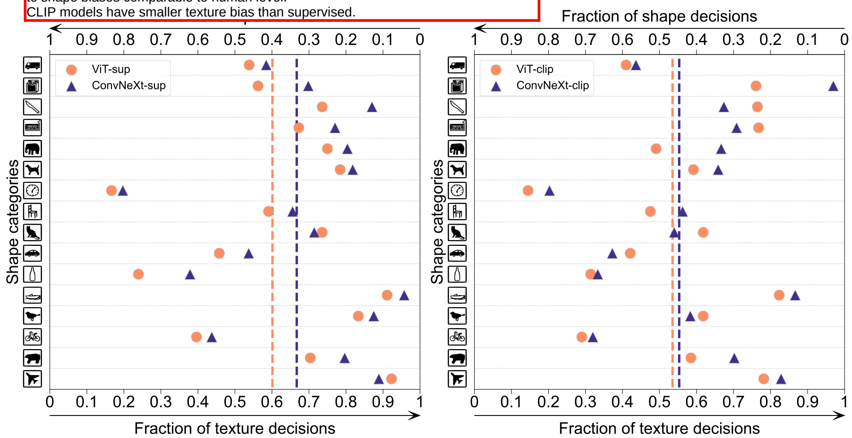


Figure 3. **Fraction of shape vs texture decisions on cue-conflict dataset.** ViT models show a higher shape bias. CLIP models are less texture-biased than their supervised counterparts. All models still have a significant fraction of texture decisions.

Figure 4. A cue-conflict image [13].

Shortcut learning in deep neural networks.

brittle shortcut features [14]. The study of shape-texture bias [13] 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 accounts for the proportion favoring the class represented by the texture. These metrics reveal whether the classifier favors shape or texture when they conflict.

The study in [13] demonstrates that ConvNets have a strong bias towards texture, as opposed to shape, which differs from human behavior. Subsequent work [36] concluded that ViT is less biased towards the texture than ConvNet by comparing the first generation of DeiT-S [50] and ResNet-50. Notably, scaling large Transformer models has led to shape biases comparable to human level [8].

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 represented with a 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 operations. 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. In [8], it has been shown that a 22B parameter ViT model can achieve 87% shape bias. In our analysis, the highest result for ViT CLIP is 46.4%, suggesting that the model size might also play an important role.

https://youtu.be/hWb-MIXKe-s?si=iOSMpHuQHHufsgcr
https://youtu.be/g4VuFAFIeE0?si=R0XXR9ZJl2RHhUE9

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 [16]. 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 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 the following way:

$$\mathrm{ECE} = \sum_{i}^{M} \frac{|B_i|}{n} \left|\mathrm{acc}(B_i) - \mathrm{conf}(B_i)\right|, \tag{1}$$

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

https://github.com/hollance/reliability-diagrams

Model calibration is also often assessed through visualizations, including reliability diagrams and confidence histograms and ECE metric.(Expected Calibration Error should be low for well calibrated models)
we present calibration metrics
on two different datasets: ImageNet-1K for in -distribution data and ImageNet-R for out-of-distribution data

CLIP models are overconfident and supervised models are slightly underconfident.
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 indistribution and out-of-distribution data
Higher ECE scores in CLIP models despite superior accuracy on ImageNet-R, indicate that training data and objectives could be more influential factors.

In the lower section 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 (3rd row). Additionally, these models show elevated accuracy levels in the initial bins of the reliability diagrams (4th row). These findings suggest that supervised models tend to be slightly underconfident on ImageNet-R
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
Supervised ConvNeXt is better calibrated than supervised ViT
For CLIP models, we find that ViT is only slightly better than ConvNeXt.

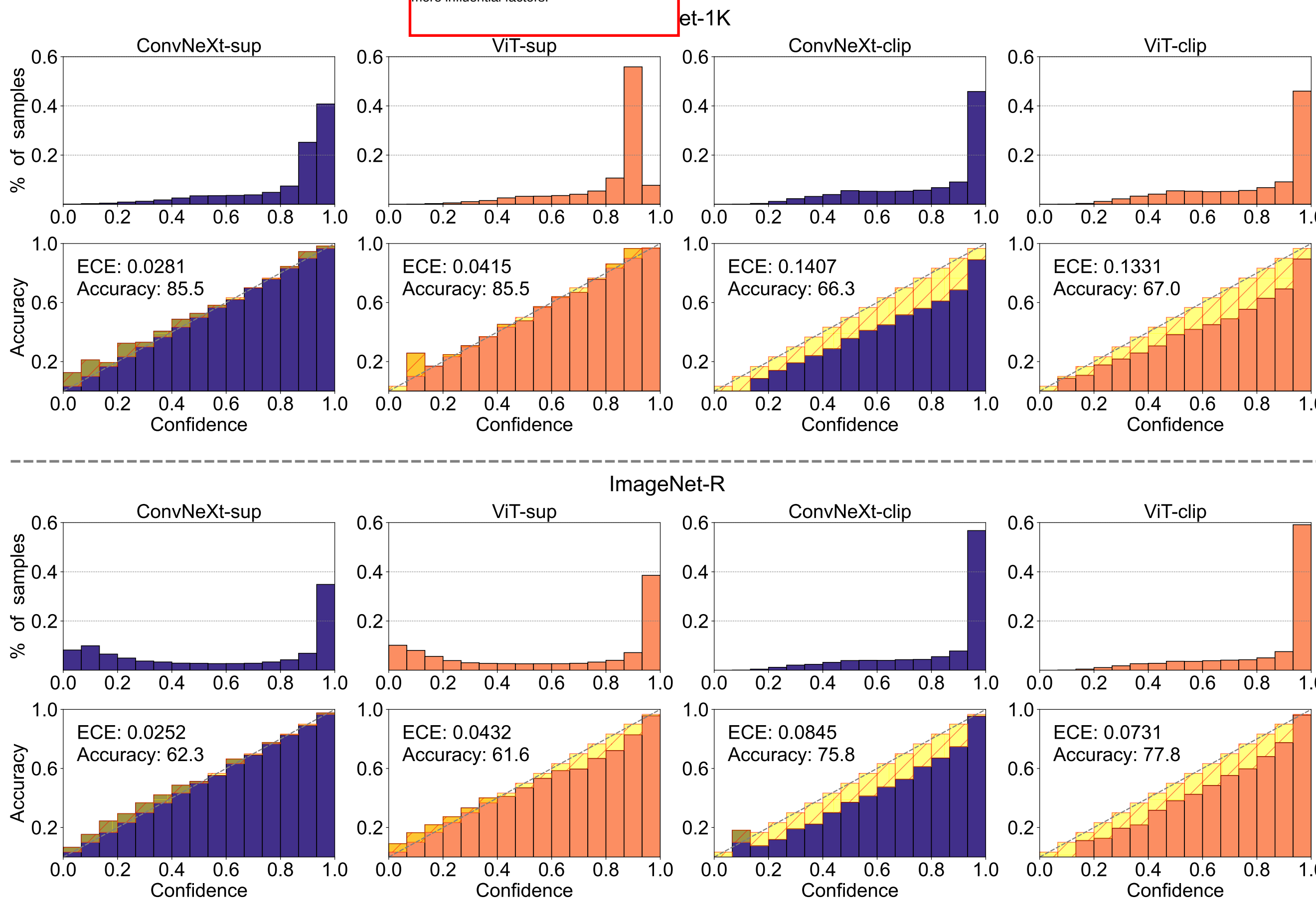


Figure 5. **Calibration results**: confidence histograms (1st and 3rd row), reliability diagrams (2nd and 4th row), and ECE metric on ImageNet-1K (top) and ImageNet-R (bottom). Supervised models have lower ECE in both cases. CLIP models have bars under diagonal and many high confidence predictions, indicating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

Revisiting the calibration of modern neural networks

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 [23] for out-of-distribution data. We select ImageNet-R as the out-of-distribution dataset because CLIP models show higher accuracy on it than supervised. In all experiments, we divide the data into $M = 15$ bins. We plot confidence histograms (1st and 3rd rows), reliability diagrams (2nd and 4th 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 [35] 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 [35] for CLIP models presumably because they use checkpoints from OpenAI [40] and we use from OpenCLIP [26]. In the lower section 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 (3rd row). Additionally, these models show elevated accuracy levels in the initial bins of the reliability diagrams (4th row). These findings suggest that supervised models tend to be slightly underconfident on ImageNet-R.

Supervised ConvNeXt is better calibrated than supervised ViT. Contrary to [35], which finds that ViTs are better calibrated than ConvNets, our experiments show that supervised ConvNeXt is better calibrated than its Transformer counterpart. This discrepancy is because [35] 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.

Supervised models are generally superior on ImageNet robustness benchmarks(Imanget distribution shifts - ImageNet-V2, ImageNet-A,ImageNet-C, ImageNet-R, ImageNet-Sketch, ImageNet-Real and ImageNet-Hard) 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 the supervised models were finetuned. ViT and ConvNeXt, on average, exhibit similar performance across both supervised and CLIP in robustness.

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

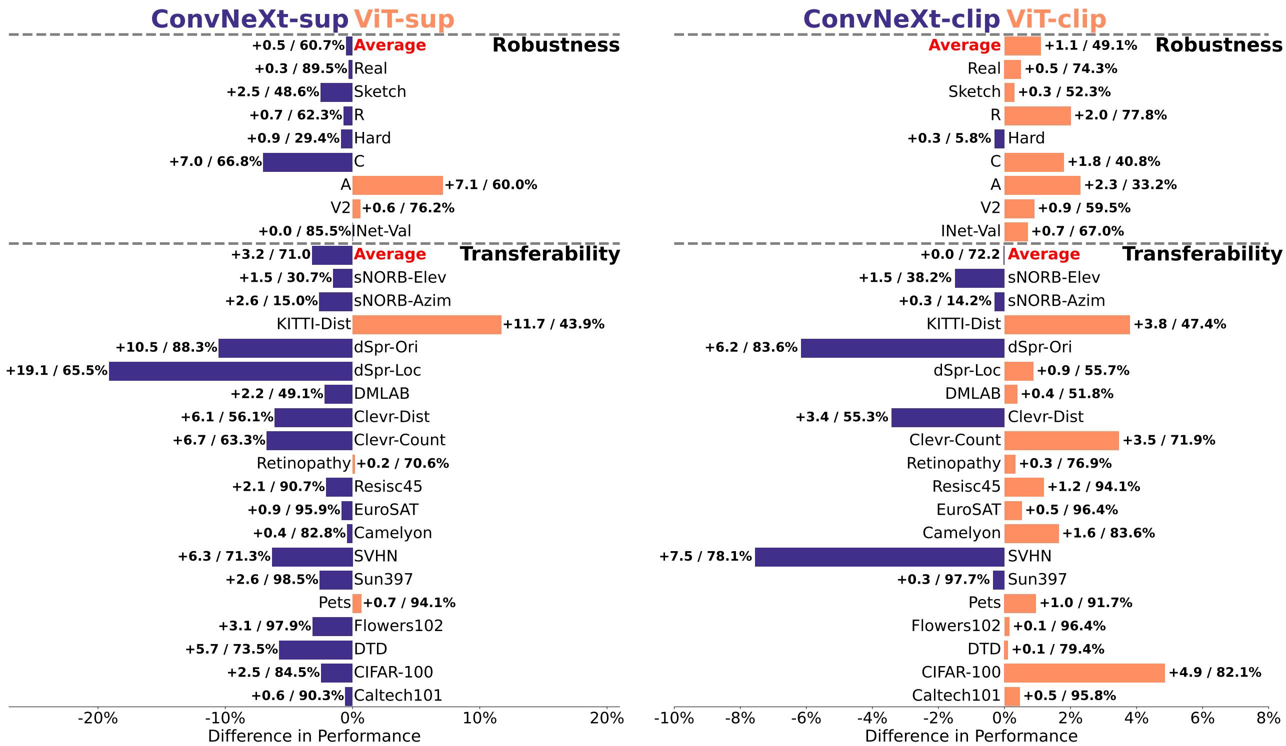


Figure 6. **Robustness (top) and transferability (bottom) results.** CLIP models excel in transferability, while supervised models are better on robustness benchmarks. In transferability, supervised ConvNeXt outperforms supervised ViT and is close to CLIP models.

Do imagenet classifiers generalize to imagenet?

3.4. Robustness

A model may excel on data from its training distribution but struggle to generalize to a distribution shift [44]. 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 benchmarks that feature many different types of natural variations and corruptions: ImageNet-V2 [44], ImageNet-A [24], ImageNet-C [22], ImageNet-R [23], ImageNet-Sketch [56], ImageNet-Real [5] and ImageNet-Hard [48]. We also provide ImageNet-1K validation accuracy for reference (INet-Val). We present the result plot in Fig. 6 (top half).

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 the supervised models were finetuned. This highlights the need for the development of new robustness benchmarks non-related to ImageNet. ViT and ConvNeXt, on average, exhibit 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 [27]. 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 evaluate how well this model performs when applied to a vastly different domain, such as medical imaging.

To assess the transferability of models, we adopted a VTAB benchmark [62]. 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 [26]. The results are shown in Fig. 6 (bottom). Transferability results on VTAB, grouped by subcategories are provided in Table 2.

ConvNeXt is better than ViT on synthetic data. Intriguingly, ConvNeXt outperforms ViT on PUG-ImageNet
for nearly all factors except Scene Light, for which all models perform poorly. This suggests
: ConvNeXt is better than
ViT on synthetic data. For CLIP models, the gap between
ConvNeXt and ViT is slightly smaller than that for supervised models, and they generally
have lower accuracy compared to supervised models. This is likely related to their
inferior accuracy on the original ImageNet.

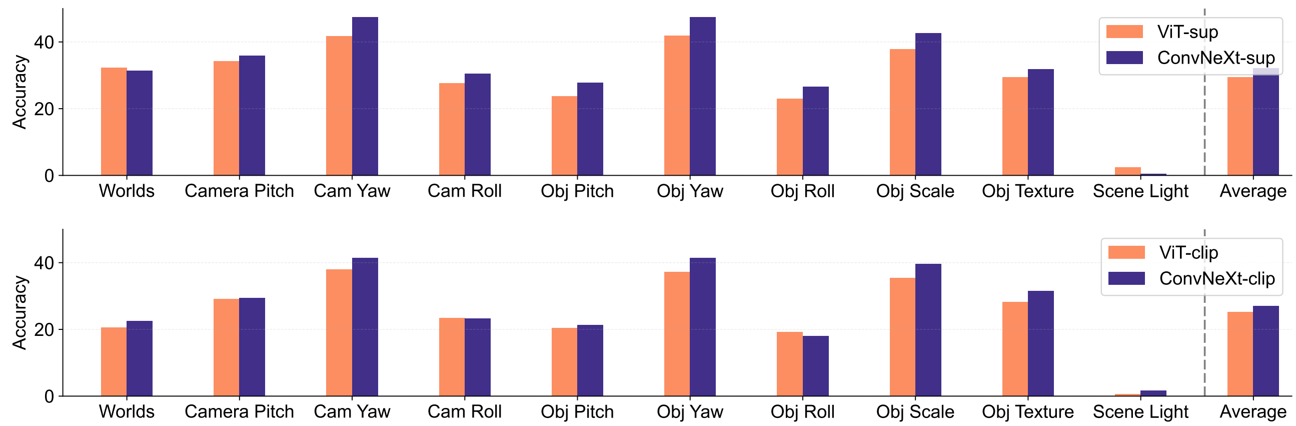


Figure 7. **Results on synthetic data from PUG-ImageNet.** Higher is better. ConvNeXt is superior on almost every factor across both supervised and CLIP.

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 models. Interestingly the performance of supervised ConvNeXt is not very far from CLIP models, both of which have the same average accuracy. For CLIP models, ViT and ConvNeXt demonstrate similar average performance, with many datasets showing a performance gap of less than 1%. CLIP models generally show better transferability on all three subgroups (Table 2) of VTAB, which is different from the robustness experiments. Their superiority can be attributed to the larger and more diverse pretraining data [42].

3.6. Synthetic Data

Synthetic data from diffusion models improves imagenet classification
Is synthetic data from generative models ready for image recognition?
Stablerep: Synthetic images from text-toimage models make strong visual representation learners.

While two previous sections focused on robustness and transferability benchmarks, these do not capture the recently emerged promising research direction of training models on synthetic data [1, 21, 49]. Unlike human-annotated data, synthetic datasets allow precise control over factors like camera angles, object positioning, and textures.

PUG-ImageNet [6] is a synthetic dataset of photorealistic images of ImageNet classes that also provides labels for a set of factors. The images are generated using a software engine that allows systematically varying factors like pose, size, texture, lighting, and background for each object. Unlike ImageNet-X, the performance on PUG-ImageNet is measured by absolute top-1 model accuracy. 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 except Scene Light, for which all models perform poorly. This suggests: ConvNeXt is better than ViT on synthetic data. For CLIP models, the gap between ConvNeXt and ViT is slightly smaller than that for supervised models, and they generally have lower accuracy compared to supervised models. This is likely related to their inferior accuracy on the original ImageNet.

3.7. Transformation Invariance

Why do deep convolutional networks generalize so poorly to small image transformations?
Making convolutional networks shiftinvariant again

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 [2, 63], it has been shown that the performance of neural networks can be highly unstable even under simple input data transformations, such as shifting an 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 crop experiments, the image is 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 (non-resized) image space and then take a crop, shifting along the longer side of the image. In resolution experiments with the ViT model, we interpolate positional embeddings to match the applied resolution.

Supervised ConvNeXt excels in transformation invariance. The results are presented in Fig . 8. We observe a very consistent trend of ConvNeXt outperforming ViT under supervised training. This trend is reversed for CLIPmodels, likely because ConvNeXt-clip was undertrained.
Overall, models are reliable to shift transformation and less robust to scale and resolution transforms. For practical applications requiring high robustness to scale, shift, and resolution transforms, our results indicate that ConvNeXt under supervised training could be the best choice.

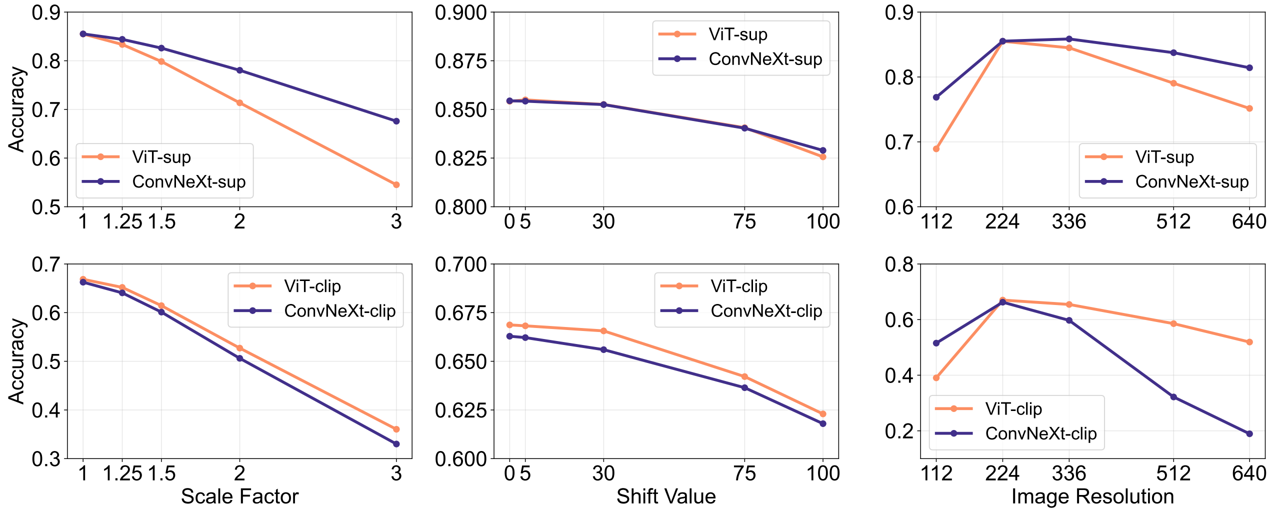


Figure 8. **Scale, shift, and resolution invariance experiments.** ConvNeXt is better than ViT under supervised training on all transformation types. All models are robust to shift transformation but experience degradation when the image scale is altered. Interestingly, supervised ConvNeXt has better performance on 336 pixel resolution.

Supervised ConvNeXt excels in transformation invariance. The results are presented in Fig. 8. We observe a very consistent trend of ConvNeXt outperforming ViT under supervised training. This trend is reversed for CLIP models, likely because ConvNeXt-clip was undertrained. Overall, models are reliable to shift transformation and less robust to scale and resolution transforms. For practical applications requiring high robustness to scale, shift, and resolution transforms, our results indicate that ConvNeXt under supervised training could be the best choice.

4. Related work

Architecture analysis. Several works compared ViTs and ConvNeXt from the perspective of internal representations [41], synthetic data [46], transferability [64] and robustness [4, 10, 39, 57]. Other studies included analysis of Transformer properties [36] and impact of neural network width and depth on learned representations [37]. In [53], ViT and CNNs were evaluated on ImageNet classification tasks, showing that Transformers are more aligned with human error patterns. While most of these works compared architectures for a single property, our work offers a comprehensive view with different training paradigms and ensures a more fair setting by selecting models with similar ImageNet accuracies.

Teaching matters: Investigating the role of supervision in vision transformers

Training objective analysis. A comprehensive analysis was conducted in [55], 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 [15, 17], aiming to find similarities and differences between the two. Two works [38, 47] focused on investigating the effect of training objective in masked image modeling and contrastive learning. In contrast to these studies that put an emphasis on self-supervised models, our study focuses on comparing image-only supervised models and image-text CLIP models.

Limitations of ImageNet. Recent research [5, 44, 52, 61] highlighted issues with the reliability and quality of ImageNet labels, suggesting they may not be good indicators of a model's ability to generalize. Two studies [28, 34] showed a strong relationship between performance on ImageNet and other datasets, although this can depend on the model's architecture and training methods. Another set of works [12, 45] emphasized that achieving high ImageNet accuracy is not a guaranteed indicator of strong performance on more diverse datasets. Current robustification training techniques were found to overfit [60] to ImageNet evaluations. In addition, ImageNet suffers from dichotomous data difficulty [33], where the majority of images are either trivially easy or impossibly hard for models to classify correctly, obscuring differences between models. Our analysis does not directly address data-related problems of ImageNet but instead studies alternative properties.

Does robustness on imagenet transfer to downstream tasks?

5. Conclusion

This study examined ConvNets and Transformers with supervised and CLIP training from multiple perspectives beyond the standard ImageNet accuracy metric. We found that each model can have its own distinct strengths. This suggests that model selection should depend on the target use cases, as standard performance measures may overlook key nuances. Additionally, many existing benchmarks are derived from ImageNet which biases the evaluation. Developing entirely new benchmarks with different data distributions will be fundamental for evaluating model performance in a broader, more real-world representative context.

For supervised models, we found the superior performance of ConvNeXt over ViT on many benchmarks. It is better calibrated, more invariant to data transformation, and has shown superior transferability and robustness. Moreover, both CLIP and supervised ConvNeXt are better than ViT on synthetic data. While CLIP models have lower ImageNet accuracy, they exhibit higher shape bias and better transferability, making them preferable in scenarios with significant domain shifts. Finally, all models make largely the same types of mistakes and struggle with texture.

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 has the strongest performance among all. In case of a serious domain shift, both of the CLIP models should provide competitive performance.

Acknowledgements

We would like to thank Sheng Zhang, Kudaibergen Abutalip, Guangyi Chen, Nurdaulet Mukhituly, Boyang Sun, Zhiqiu Xu, and Nurbek Tastan for their valuable feedback.

References

[1] Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J Fleet. Synthetic data from diffusion models improves imagenet classification. *arXiv preprint arXiv:2304.08466*, 2023.

[2] Aharon Azulay and Yair Weiss. Why do deep convolutional networks generalize so poorly to small image transformations? *arXiv preprint arXiv:1805.12177*, 2018.

[3] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[4] Yutong Bai, Jieru Mei, Alan L Yuille, and Cihang Xie. Are transformers more robust than cnns? In *NeurIPS*, 2021.

[5] Lucas Beyer, Olivier J Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with imagenet? *arXiv preprint arXiv:2006.07159*, 2020.

[6] Florian Bordes, Shashank Shekhar, Mark Ibrahim, Diane Bouchacourt, Pascal Vincent, and Ari S Morcos. Pug: Photorealistic and semantically controllable synthetic data for representation learning. *arXiv preprint arXiv:2308.03977*, 2023.

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020.

[8] Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, et al. Scaling vision transformers to 22 billion parameters. In *ICML*, 2023.

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[10] 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 *CVPR*, 2021.

[11] 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. *arXiv preprint arXiv:2010.11929*, 2020.

[12] Alex Fang, Simon Kornblith, and Ludwig Schmidt. Does progress on imagenet transfer to real-world datasets? *arXiv preprint arXiv:2301.04644*, 2023.

[13] 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.

[14] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2020.

[15] Tom George Grigg, Dan Busbridge, Jason Ramapuram, and Russ Webb. Do self-supervised and supervised methods learn similar visual representations? *arXiv preprint arXiv:2110.00528*, 2021.

[16] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *ICML*, 2017.

[17] Matthew Gwilliam and Abhinav Shrivastava. Beyond supervised vs. unsupervised: Representative benchmarking and analysis of image representation learning. In *CVPR*, 2022.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[19] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020.

[20] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022.

[21] Ruifei He, Shuyang Sun, Xin Yu, Chuhui Xue, Wenqing Zhang, Philip Torr, Song Bai, and Xiaojuan Qi. Is synthetic data from generative models ready for image recognition? *arXiv preprint arXiv:2210.07574*, 2022.

[22] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

[23] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *CVPR*, 2021.

[24] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *CVPR*, 2021.

[25] Badr Youbi Idrissi, Diane Bouchacourt, Randall Balestriero, Ivan Evtimov, Caner Hazirbas, Nicolas Ballas, Pascal Vincent, Michal Drozdzal, David Lopez-Paz, and Mark Ibrahim. Imagenet-x: Understanding model mistakes with factor of variation annotations. *arXiv preprint arXiv:2211.01866*, 2022.

[26] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021.

[27] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In *ECCV*, 2020.

[28] Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better imagenet models transfer better? In *CVPR*, 2019.

[29] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. 2012.

[30] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998.

[31] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *CVPR*, 2022.

[32] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 2022.

[33] Kristof Meding, Luca M Schulze Buschoff, Robert Geirhos, and Felix A Wichmann. Imagenet suffers from dichotomous data difficulty. In *NeurIPS 2021 Workshop on ImageNet: Past, Present, and Future*, 2021.

[34] 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 *ICML*, 2021.

[35] Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. Revisiting the calibration of modern neural networks. *NeurIPS*, 2021.

[36] Muhammad Muzammal Naseer, Kanchana Ranasinghe, Salman H Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. In *NeurIPS*, 2021.

[37] Thao Nguyen, Maithra Raghu, and Simon Kornblith. 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.

[38] Namuk Park, Wonjae Kim, Byeongho Heo, Taekyung Kim, and Sangdoo Yun. What do self-supervised vision transformers learn? *arXiv preprint arXiv:2305.00729*, 2023.

[39] Francesco Pinto, Philip HS Torr, and Puneet K. Dokania. An impartial take to the cnn vs transformer robustness contest. In *ECCV*, 2022.

[40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[41] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *NeurIPS*, 2021.

[42] Vivek Ramanujan, Thao Nguyen, Sewoong Oh, Ludwig Schmidt, and Ali Farhadi. On the connection between pre-training data diversity and fine-tuning robustness. *arXiv preprint arXiv:2307.12532*, 2023.

[43] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022.

[44] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *ICML*, 2019.

[45] Megan Richards, Polina Kirichenko, Diane Bouchacourt, and Mark Ibrahim. Does progress on object recognition benchmarks improve real-world generalization? *arXiv preprint arXiv:2307.13136*, 2023.

[46] Nataniel Ruiz, Sarah Bargal, Cihang Xie, Kate Saenko, and Stan Sclaroff. Finding differences between transformers and convnets using counterfactual simulation testing. *NeurIPS*, 2022.

[47] Shashank Shekhar, Florian Bordes, Pascal Vincent, and Ari Morcos. Objectives matter: Understanding the impact of self-supervised objectives on vision transformer representations. *arXiv preprint arXiv:2304.13089*, 2023.

[48] Mohammad Reza Taesiri, Giang Nguyen, Sarra Habchi, Cor-Paul Bezemer, and Anh Nguyen. Imagenet-hard: The hardest images remaining from a study of the power of zoom and spatial biases in image classification.

[49] Yonglong Tian, Lijie Fan, Phillip Isola, Huiwen Chang, and Dilip Krishnan. Stablerep: Synthetic images from text-to-image models make strong visual representation learners. *arXiv preprint arXiv:2306.00984*, 2023.

[50] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *ICML*, 2021.

[51] Hugo Touvron, Matthieu Cord, and Hervé Jégou. Deit iii: Revenge of the vit. In *ECCV*, 2022.

[52] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. From ImageNet to image classification: Contextualizing progress on benchmarks. In *ICML*, 2020.

[53] Shikhar Tuli, Ishita Dasgupta, Erin Grant, and Thomas L Griffiths. Are convolutional neural networks or transformers more like human vision? *arXiv preprint arXiv:2105.07197*, 2021.

[54] Yael Vinker, Yuval Alaluf, Daniel Cohen-Or, and Ariel Shamir. Clipascene: Scene sketching with different types and levels of abstraction. In *ICCV*, 2023.

[55] Matthew Walmer, Saksham Suri, Kamal Gupta, and Abhinav Shrivastava. Teaching matters: Investigating the role of supervision in vision transformers. In *CVPR*, 2023.

[56] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *NeurIPS*, 2019.

[57] Zeyu Wang, Yutong Bai, Yuyin Zhou, and Cihang Xie. Can cnns be more robust than transformers? *arXiv preprint arXiv:2206.03452*, 2022.

[58] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *CVPR*, 2023.

[59] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *CVPR*, 2022.

[60] Yutaro Yamada and Mayu Otani. Does robustness on imagenet transfer to downstream tasks? In *CVPR*, 2022.

[61] Sangdoo Yun, Seong Joon Oh, Byeongho Heo, Dongyoon Han, Junsuk Choe, and Sanghyuk Chun. Re-labeling imagenet: from single to multi-labels, from global to localized labels. In *CVPR*, 2021.

[62] 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.

[63] Richard Zhang. Making convolutional networks shift-invariant again. In *ICML*, 2019.

[64] Hong-Yu Zhou, Chixiang Lu, Sibei Yang, and Yizhou Yu. Convnets vs. transformers: Whose visual representations are more transferable? In *CVPR*, 2021.