A DISTRIBUTION ESTIMATION

A.1 PROBLEM FORMULATION

Given an unknown ground truth distribution:

$$\mathbf{P} = \text{Unknown}(\mu, \Sigma) \tag{1}$$

where $\mu \in \mathbb{R}^p$ and $\Sigma \in \mathbb{R}^{p \times p}$.

All the samples in our study are sampled from this distribution.

We use $\mathbf{X}_k$ to denote the k^{th} dataset, with n_k samples, and we use $\mathbf{x}_{k,i}$ to denote the i^{th} sample in it.

We aim to consider the estimation of μ from two different models. The conventional smaller model which operates on only one dataset, and WLOG, we assume the smaller model works on $\mathbf{X}_0$; and the bigger, zoo of CLIP-style models, which operates on a collection of datasets, we say it works on m datasets, i.e., $\{\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \ldots, \mathbf{X}_m\}$, we will compare $\mathbb{E}[\widehat{\mu_0} - \mu]$ and $\mathbb{E}[\widehat{\mu_{\text{CLIP}}} - \mu]$, $\text{VAR}(\widehat{\mu_0})$ and $\text{VAR}(\widehat{\mu_{\text{CLIP}}})$, $\mathbb{E}[\widehat{\Sigma_0} - \Sigma]$ and $\mathbb{E}[\widehat{\Sigma_{\text{CLIP}}} - \Sigma]$ and $\text{VAR}(\widehat{\Sigma_0})$ and $\text{VAR}(\widehat{\Sigma_{\text{CLIP}}})$.

Assumption I Due to dataset collection bias, we assume that, while all the data are sampled with the fixed distribution above, the bias of dataset collection will introduce a bias in the estimation of the true parameter μ, therefore

$$\widehat{\mu_i} = \mu + \epsilon_i \tag{2}$$

where

$$\widehat{\mu_i} := \frac{1}{n_i} \sum_{j}^{n_i} \mathbf{x}_{i,j} \tag{3}$$

and

$$\epsilon_i \sim N(\mathbf{0}, \mathbf{I}) \tag{4}$$

Assumption II Due to dataset collection bias, we assume that, while all the data are sampled with the fixed distribution above, the bias of dataset collection will introduce a bias in the estimation of the true parameter Σ, therefore

$$\widehat{\Sigma_i} = \epsilon_i^{'} \Sigma \tag{5}$$

where

$$\widehat{\Sigma_i} := \frac{1}{n_i} \sum_{j}^{n_i} [(\mathbf{x}_{i,j} - \widehat{\mu_i})^T (\mathbf{x}_{i,j} - \widehat{\mu_i})] \tag{6}$$

and

$$\epsilon_i^{'} \sim \text{Exp}(\mathbf{1}), \tag{7}$$

Proposition A.1. *Under Assumptions I and II, we have estimators*

$$\mathbb{E}[\widehat{\mu_{\text{CLIP}}} - \mu] = \mathbb{E}[\widehat{\mu_0} - \mu], \qquad \mathbb{E}[\widehat{\Sigma_{\text{CLIP}}} - \Sigma] = \mathbb{E}[\widehat{\Sigma_0} - \Sigma]$$
$$\text{VAR}(\widehat{\mu_{\text{CLIP}}}) \leq \text{VAR}(\widehat{\mu_0}), \qquad \text{VAR}(\widehat{\Sigma_{\text{CLIP}}}) \leq \text{VAR}(\widehat{\Sigma_0})$$

where $\leq$ holds element-wise.

Proof. **Estimation of μ.** Under Assumptions I and II, we have

$$\widehat{\mu_0} = \mu + \epsilon_0 \tag{8}$$

We can obtain $\mathbb{E}[\widehat{\mu_0} - \mu]$ and $\text{VAR}(\widehat{\mu_0})$ by marginalizing out the randomness introduced by ϵ:

$$\mathbb{E}[\widehat{\mu_0} - \mu] = \mathbb{E}[\mu + \epsilon_0 - \mu] = \mathbb{E}[\epsilon_0] = \mathbf{0}. \tag{9}$$

$$\begin{aligned}\text{VAR}(\widehat{\mu_0}) &= \mathbb{E}[\widehat{\mu_0}^2] - \mathbb{E}^2[\widehat{\mu_0}] \\ &= \mathbb{E}[(\mu+\epsilon_0)^2] - \mathbb{E}^2[(\mu+\epsilon_0)] \\ &= \mathbb{E}[\mu^2 + 2\mu\epsilon_0 + \epsilon_0^2] - (\mu + \mathbb{E}[\epsilon_0])^2 \\ &= \mathbb{E}[\epsilon_0^2] - \mathbb{E}^2[\epsilon_0] \\ &= \text{VAR}(\epsilon_0) \\ &= \mathbf{I} \end{aligned} \tag{10}$$

For $\mathbb{E}[\widehat{\mu_{\text{CLIP}}} - \mu]$ and $\text{VAR}(\widehat{\mu_{\text{CLIP}}})$, we have:

$$\mathbb{E}[\widehat{\mu_{\text{CLIP}}} - \mu] = \mathbb{E}[\frac{\sum_i^m \epsilon_i n_i}{\sum_i^m n_i}] = \frac{\sum_i^m \mathbb{E}[\epsilon_i] n_i}{\sum_i^m n_i} = \mathbf{0}. \tag{11}$$

and

$$\begin{aligned}\text{VAR}(\widehat{\mu_{\text{CLIP}}}) &= \mathbb{E}[\widehat{\mu_{\text{CLIP}}}^2] - \mathbb{E}^2[\widehat{\mu_{\text{CLIP}}}] \\ &= \mathbb{E}[(\mu+\epsilon_{\text{CLIP}})^2] - \mathbb{E}^2[(\mu+\epsilon_{\text{CLIP}})] \\ &= \mathbb{E}[\mu^2 + 2\mu\epsilon_{\text{CLIP}} + \epsilon_{\text{CLIP}}^2] - (\mu + \mathbb{E}[\epsilon_{\text{CLIP}}])^2 \\ &= \mathbb{E}[\epsilon_{\text{CLIP}}^2] - \mathbb{E}^2[\epsilon_{\text{CLIP}}] \end{aligned} \tag{12}$$

Since $\mathbb{E}[\epsilon_{\text{CLIP}}] = \mathbb{E}[\widehat{\mu_{\text{CLIP}}} - \mu] = 0$, we have:

$$\text{VAR}(\widehat{\mu_{\text{CLIP}}}) = \mathbb{E}[\epsilon_{\text{CLIP}}^2] = \mathbb{E}[(\frac{\sum_i^m \epsilon_i n_i}{\sum_i^m n_i})^2] \tag{13}$$

When we expand the square of sum, we will get the many squared terms (which are left finally) and many more that involves at least one $\mathbb{E}[\epsilon_i]\mathbf{z}$, where $\mathbf{z}$ can be any arbitrary stuff, and then since $\mathbb{E}[\epsilon_i] = \mathbf{0}$, $\mathbf{z}$ won't matter. Therefore, we have:

$$\text{VAR}(\widehat{\mu_{\text{CLIP}}}) = \mathbb{E}[(\frac{\sum_i^m \epsilon_i n_i}{\sum_i^m n_i})^2] = \frac{\sum_i^m n_i^2 \mathbb{E}[\epsilon_i^2]}{(\sum_i^m n_i)^2} \tag{14}$$

Since $n_i \geq 1$ for $i = 1, 2, ..., m$, we have $\sum_i^m n_i^2 \leq (\sum_i^m n_i)^2$.

Therefore,

$$\text{VAR}(\widehat{\mu_{\text{CLIP}}}) = \frac{\sum_i^m n_i^2 \mathbb{E}[\epsilon_i^2]}{(\sum_i^m n_i)^2} \leq \frac{(\sum_i^m n_i)^2 \mathbb{E}[\epsilon_i^2]}{(\sum_i^m n_i)^2} = \mathbb{E}[\epsilon_i^2] = \mathbf{I} \tag{15}$$

Estimation of Σ. We can obtain $\mathbb{E}[\widehat{\Sigma_0} - \Sigma]$ and $\text{VAR}(\widehat{\Sigma_0})$ by marginalizing out the randomness introduced by $\epsilon^{'}$:

$$\mathbb{E}[\widehat{\Sigma_0} - \Sigma] = \mathbb{E}[\epsilon_0^{'}\Sigma - \Sigma] = \mathbb{E}[(\epsilon_0^{'} - 1)\Sigma] = \mathbb{E}[\epsilon_0^{'} - 1]\mathbb{E}[\Sigma] = \mathbf{0}. \tag{16}$$

$$\begin{aligned}\text{VAR}(\widehat{\Sigma_0}) &= \mathbb{E}[\widehat{\Sigma_0}^2] - \mathbb{E}^2[\widehat{\Sigma_0}] \\ &= \mathbb{E}[(\epsilon_0^{'}\Sigma)^2] - \mathbb{E}^2[\epsilon_0^{'}\Sigma] \\ &= \mathbb{E}[\epsilon_0^{'2}\Sigma^2] - \mathbb{E}^2[\epsilon_0^{'}]\mathbb{E}^2[\Sigma] \\ &= \mathbb{E}[\epsilon_0^{'2}]\mathbb{E}[\Sigma^2] - \mathbb{E}^2[\epsilon_0^{'}]\mathbb{E}^2[\Sigma] \\ &= 2\mathbb{E}[\Sigma^2] - \mathbb{E}^2[\Sigma] \\ &= 2\Sigma^2 - \Sigma^2 \\ &= \Sigma^2 \end{aligned} \tag{17}$$

For $\mathbb{E}[\widehat{\Sigma_{\text{CLIP}}} - \Sigma]$ and $\text{VAR}(\widehat{\Sigma_{\text{CLIP}}})$, we have:

$$\mathbb{E}[\widehat{\Sigma_{\text{CLIP}}} - \Sigma] = \mathbb{E}[\epsilon_{\text{CLIP}}^{'}\Sigma - \Sigma] = \mathbb{E}[(\epsilon_{\text{CLIP}}^{'} - 1)\Sigma] = \mathbb{E}[\epsilon_{\text{CLIP}}^{'} - 1]\mathbb{E}[\Sigma] = \mathbf{0}. \tag{18}$$

$$\begin{aligned}
\text{VAR}(\widehat{\Sigma_{\text{CLIP}}}) &= \mathbb{E}[\widehat{\Sigma_{\text{CLIP}}}^2] - \mathbb{E}^2[\widehat{\Sigma_{\text{CLIP}}}] \\
&= \mathbb{E}[(\epsilon_{\text{CLIP}}^{'}\Sigma)^2] - \mathbb{E}^2[\epsilon_{\text{CLIP}}^{'}\Sigma] \\
&= \mathbb{E}[\epsilon_{\text{CLIP}}^{'2}\Sigma^2] - \mathbb{E}^2[\epsilon_{\text{CLIP}}^{'}]\mathbb{E}^2[\Sigma] \\
&= \mathbb{E}[\epsilon_{\text{CLIP}}^{'2}]\mathbb{E}[\Sigma^2] - \mathbb{E}^2[\Sigma]
\end{aligned} \tag{19}$$

Consider that

$$\widehat{\Sigma_{\text{CLIP}}} = \frac{\Sigma_i^m \Sigma_j^{n_i}(\mathbf{x}_{i,j} - \widehat{\mu_{\text{CLIP}}})^2}{\Sigma_i^m n_i} = \frac{\Sigma_i^m \epsilon_{\text{CLIP}}^2 n_i}{\Sigma_i^m n_i} = \epsilon_{\text{CLIP}}^{'}\Sigma \tag{20}$$

We will have:

$$\epsilon_{\text{CLIP}}^{'} = \frac{\epsilon_{\text{CLIP}}^2}{\Sigma} \tag{21}$$

Thus, we have $\epsilon_{\text{CLIP}}^{'2} = \frac{\epsilon_{\text{CLIP}}^4}{\Sigma^2}$. Next, we will compute $\mathbb{E}[\epsilon_{\text{CLIP}}^{'2}]$ as follows:

$$\begin{aligned}
\mathbb{E}[\epsilon_{\text{CLIP}}^{'2}] &= \mathbb{E}[\frac{\epsilon_{\text{CLIP}}^4}{\Sigma^2}] \\
&= \frac{\mathbb{E}[\epsilon_{\text{CLIP}}^4]}{\Sigma^2}
\end{aligned} \tag{22}$$

By definition, we have:

$$\epsilon_{\text{CLIP}} = \frac{\Sigma_i^m \Sigma_j^{n_i}(\mathbf{x}_{i,j} - \widehat{\mu_i})}{\Sigma_i^m n_i} \tag{23}$$

Therefore,

$$\epsilon_{\text{CLIP}}^2 = \frac{(\Sigma_i^m \Sigma_j^{n_i}(\mathbf{x}_{i,j} - \widehat{\mu_i}))^2}{(\Sigma_i^m n_i)^2} \tag{24}$$

As the value of $x_{i,j} - \widehat{\mu_i}$ can be either positive or negative, we have:

$$\epsilon_{\text{CLIP}}^2 \leq \frac{\Sigma_i^m \Sigma_j^{n_i}(\mathbf{x}_{i,j} - \widehat{\mu_i})^2}{\Sigma_i^m n_i} \frac{1}{\Sigma_i^m n_i} \tag{25}$$

Since both $(x_{i,j} - \widehat{\mu_i})^2$ and n_i are positive values, we further have:

$$\epsilon_{\text{CLIP}}^2 \leq \Sigma_i^m \frac{\Sigma_j^{n_i}(\mathbf{x}_{i,j} - \widehat{\mu_i})^2)}{n_i} \frac{1}{\Sigma_i^m n_i} = \frac{\Sigma_i^m \widehat{\Sigma_i}}{\Sigma_i^m n_i} = \frac{\Sigma_i^m \epsilon_i^{'}\Sigma}{\Sigma_i^m n_i} \tag{26}$$

Thus, we can obtain

$$\epsilon_{\text{CLIP}}^4 \leq \frac{(\Sigma_i^m \epsilon_i^{'}\Sigma)^2}{(\Sigma_i^m n_i)^2} = \frac{(\Sigma_i^m \epsilon_i^{'})^2\Sigma^2}{(\Sigma_i^m n_i)^2} \tag{27}$$

Therefore, we have:

$$\mathbb{E}[\epsilon_{\text{CLIP}}^4] \leq \mathbb{E}[\frac{(\Sigma_i^m \epsilon_i^{'})^2\Sigma^2}{(\Sigma_i^m n_i)^2}] = \frac{\mathbb{E}[(\Sigma_i^m \epsilon_i^{'})^2]\Sigma^2}{(\Sigma_i^m n_i)^2} \tag{28}$$

By Assumption A.1, $\epsilon_i^{'} \sim \text{Exp}(\mathbf{1})$, we have $\mathbb{E}[\epsilon_i^{'}] = 1$ and $\text{VAR}(\epsilon_i^{'}) = 1$.

Since $\epsilon_i^{'}$ are independent with each other, we have:

$$\begin{aligned}
\mathbb{E}[(\Sigma_i^m \epsilon_i^{'})^2] &= \text{VAR}(\Sigma_i^m \epsilon_i^{'}) + \mathbb{E}^2[\Sigma_i^m \epsilon_i^{'}] \\
&= \Sigma_i^m \text{VAR}(\epsilon_i^{'}) + (\Sigma_i^m \mathbb{E}[\epsilon_i^{'}])^2 \\
&= m + m^2
\end{aligned} \tag{29}$$

Substituting Eq. 29 into Eq. 28, we have:

$$\mathbb{E}[\epsilon_{\text{CLIP}}^4] \leq \frac{m+m^2}{(\Sigma_i^m n_i)^2}\Sigma^2 \tag{30}$$

Since $n_i \geq 1$ for $i = 1, 2, ..., m$, we have $\Sigma_i^m n_i \geq m$ and $(\Sigma_i^m n_i)^2 \geq m^2$.

Since $m \geq 1$, we have: $m^2 \geq m$.

Therefore,

$$\mathbb{E}[\epsilon_{\text{CLIP}}^4] \leq \frac{m+m^2}{m^2}\Sigma^2 \leq \frac{2m^2}{m^2}\Sigma^2 = 2\Sigma^2 \tag{31}$$

Substituting Eq. 31 into Eq. 22, we have:

$$\mathbb{E}[\epsilon_{\text{CLIP}}^{'2}] \leq \frac{2\Sigma^2}{\Sigma^2} = 2 \tag{32}$$

Substituting Eq. 32 into Eq. 19, we have:

$$\text{VAR}(\widehat{\Sigma_{\text{CLIP}}}) = \mathbb{E}[\epsilon_{\text{CLIP}}^{'2}]\mathbb{E}[\Sigma^2] - \mathbb{E}^2[\Sigma] \leq 2\mathbb{E}[\Sigma^2] - \mathbb{E}^2[\Sigma] = \mathbb{E}[\Sigma] = \Sigma \tag{33}$$

We summarize the above results as follows: For conventional fixed dataset estimators, we have:

$$\begin{aligned}
\mathbb{E}[\widehat{\mu_0} - \mu] &= \mathbf{0} \\
\text{VAR}(\widehat{\mu_0}) &= \mathbf{I} \\
\mathbb{E}[\widehat{\Sigma_0} - \Sigma] &= \mathbf{0} \\
\text{VAR}(\widehat{\Sigma_0}) &= \Sigma^2
\end{aligned}$$

For CLIP-style estimators, we have:

$$\begin{aligned}
\mathbb{E}[\widehat{\mu_{\text{CLIP}}} - \mu] &= \mathbf{0} \\
\text{VAR}(\widehat{\mu_{\text{CLIP}}}) &\leq \mathbf{I} \\
\mathbb{E}[\widehat{\Sigma_{\text{CLIP}}} - \Sigma] &= \mathbf{0} \\
\text{VAR}(\widehat{\Sigma_{\text{CLIP}}}) &\leq \Sigma,
\end{aligned}$$

where $\leq$ holds element-wise. □

The results show that, both conventional estimator and zoo of CLIP-style estimator can recover the true μ, Σ of the unknown distribution, but zoo of CLIP-style estimator will have a lower variance, which is more stable to accomplish the task. This conclusion holds for any distributions.

With these theoretical evidence, we kindly argue that biased towards the zoo of CLIP-style models is better than biased on conventional fixed datasets. In addition, recent advances in incorporating the foundation model into various tasks (Liu et al., 2023; Zhang et al., 2023; Bose et al., 2023) also reveals that the community has utilized the foundation model on a large scale and pays little attention on these biases.

B Notes on the Experimental Setup

B.1 Notes on Models

Note that we only re-evaluate existing model checkpoints, and hence do not perform any hyperparameter tuning for evaluated models. Our model evaluations are done on 8 NVIDIA V100 GPUs. With our Sparsified VQGAN model, our method is also feasible to work with a small amount of GPU resources. As shown in Appendix I, the proposed protocol can work on a single NVIDIA V100 GPU efficiently.

B.2 Hyperparameter Tuning

Our method is generally parameter-free except for the computation budget and perturbation step size. In our experiments, the computation budget is the maximum iteration number of Sparse VQGAN. We consider the predefined value to be 50, as it guarantees the degree of perturbation with acceptable time costs. We provide the experiment for step size configuration in Section 4.5.

C In-depth Analysis on the Transformer Family

In Table 1, we notice a large difference between the methods in the proposed FMR metric, even within the transformer family. After checking the distribution of misclassified perturbed images of different models, we find that these images are rather random and do not reveal any obvious "weak classes". One possible reason for this phenomenon may due to their internal architecture that are related to the self-attention (SA) mechanism. Many current Vision Transformer architectures adopt a multi-head self-attention (MHSA) design where each head tends to focus on different object components. In some sense, MHSA can be interpreted as a mixture of information bottlenecks (IB) where the stacked SA modules in Vision Transformers can be broadly regarded as an iterative repeat of the IB process which promotes grouping and noise filtering. More details of the connection between the SA and IB can be found in ((Zhou et al., 2022a), Sec.2.3). As revealed in (Zhou et al., 2022a), having more heads leads to improved expressivity and robustness. But the reduced channel number per head also causes decreased clean accuracy. The best trade-off is achieved with 32 channels per head.

Table 4 illustrates the head number configurations of various models employed in our experiment.

Table 4: Details of head numbers configurations.)

Model	Head Number
ViT	12
DeiT	12
Twins	(3,6,12,24)
Visformer	6
Swin	(4,8,16,32)

Swin Transformer exhibits the highest number of heads among them. Despite its suboptimal accuracy on the standard dataset, it achieves the best FMR. This corroborates the finding in (Zhou et al., 2022a) that increased head numbers enhance expressivity and robustness, albeit at the expense of clean accuracy.

To further verify the impact of head numbers, we trained Swin Transformer with varying head configurations and obtained the following results in Table 5.

Table 5: The performance of Swin Transformer with different head number configurations. We find that increased heads enhance expressivity and robustness.)

#Params	Head Number	SA	FMR
88M	(2,4,8,16)	80.82	64.85
88M	(3,6,12,24)	81.98	67.48
88M	(4,8,16,32)	81.67	69.73
88M	(5,10,20,40)	81.05	69.97

With comparable numbers of parameters, we observe that their accuracies on the standard dataset are relatively similar. With the augmentation of head numbers, the FMR value also escalates, which validates our hypothesis that increased heads enhance expressivity and robustness.

D Transferability of Generated Images

We first study whether our generated images are model-specific, since the generation of the images involves the gradient of the original model. We train several architectures, namely EfficientNet (Tan

& Le, 2019), MobileNet (Howard et al., 2017), SimpleDLA (Yu et al., 2018), VGG19 (Simonyan & Zisserman, 2014), PreActResNet (He et al., 2016b), GoogLeNet (Szegedy et al., 2015), and DenseNet121 (Huang et al., 2017) and test these models with the images that generated when testing ResNet. We also train another ResNet following the same procedure to check the transferability across different runs in one architecture.

Transferability of the generated images. Table 6 shows a reasonable transferability of the generated images as the FMR are all lower than the SA, although we can also observe an improvement over the FMR when tested in the new models. These results suggest that our method of generating images can be potentially used in a broader scope: we can also leverage the method to generate a static set of images and set a benchmark dataset to help the development of robustness methods.

Table 6: Performances of transferability.

Model	SA	PA	FMR
ResNet	95.38	51.67	54.17
ResNet	94.67	56.09	59.25
DenseNet	94.26	60.48	64.17
SimpleDLA	92.25	61.03	66.16
GoogLeNet	92.06	61.10	66.38
PreActResNet	90.91	61.14	67.25
EfficientNet	91.37	62.57	68.48
MobileNet	91.63	62.97	68.72
VGG	93.54	66.01	70.57

Reliability of the FMR metric. Moreover, these results contribute to the validation of the reliability of the FMR metric: given that each model's FMR gets computed using a different test set, it is not clear why FMR would be a reliable metric that can be used to compare two models. In this experiment, however, the models are tested using the same fixed test set that was initially generated during the evaluation on ResNet. Remarkably, the strong correlation observed between FMR and PA at the fixed test sets lends credence to the reliability of the FMR metric.

New findings. In addition, our results might potentially help mitigate a debate on whether more accurate architectures are naturally more robust: on one hand, we have results showing that more accurate architectures indeed lead to better empirical performances on certain (usually fixed) robustness benchmarks (Rozsa et al., 2016; Hendrycks & Dietterich, 2019); while on the other hand, some counterpoints suggest the higher robustness numerical performances are only because these models capture more non-robust features that also happen exist in the fixed benchmarks (Tsipras et al., 2018; Wang et al., 2020b; Taori et al., 2020). Table 6 show some examples to support the latter argument: in particular, we notice that VGG, while ranked in the middle of the accuracy ladder, interestingly stands out when tested with generated images. These results continue to support our argument that a dynamic robustness test scenario can help reveal more properties of the model.

E Initiating with Adversarial Attacked Images

Since our method using the gradient of the evaluated model reminds readers about the gradient-based attack methods in adversarial robustness literature, we test whether initiating the perturbation process with an adversarial example will further degrade the accuracy.

Table 7: Results on whether initiating with adversarial images ($\epsilon = 0.003$).

Data	SA	FMR
regular	95.38	57.80
w. FGSM	95.30	65.79

We first generate the images with FGSM attack (Goodfellow et al., 2015). Table 7 shows that initiating with the FGSM adversarial examples barely affect the FMR, which is probably because the major style-wise perturbation will erase the imperceptible perturbations the adversarial examples introduce.

F Adversarially Robust Models

With evidence suggesting the adversarially robust models are considered more human perceptually aligned (Engstrom et al., 2019; Zhang & Zhu, 2019; Wang et al., 2020b), we compare the vanilla model to a model trained by PGD (Madry et al., 2017) (ℓ_∞ norm smaller than 0.03).

As shown in Table 8, adversarially trained model and vanillaly trained model indeed process the data differently: the transferability of the generated images between these two regimes can barely hold. In particular, the PGD model can almost maintain its performances when tested with the images generated by the vanilla model.

However, despite the differences, the PGD model's robustness weak spots are exposed to a similar degree with the vanilla model by our test system: the FMR of the vanilla model and the PGD model are only 57.79 and 66.18, respectively. We believe this result can further help advocate our belief that the robustness test needs to be a dynamic process generating images conditioning on the model to test, and thus further help validate the importance of our contribution.

Table 8: Performances comparison with vanilla model and PGD trained model.

Data	Model	SA	FMR
Van.	Van.	95.38	57.79
	PGD	85.70	95.96
PGD	Van.	95.38	62.41
	PGD	85.70	66.18

G Augmentation through Static Adversarial Training

Intuitively, inspired by the success of adversarial training (Madry et al., 2017) in defending models against adversarial attacks, a natural method to improve the empirical performances under our new test protocol is to augment the training data with perturbed training images generated by the same process. We aim to validate the effectiveness of this method here.

However, the computational load of generation process is not ideal to serve the standard adversarial training strategy, and we can only have one copy of the perturbed training samples. Fortunately, we notice that some recent advances in training with data augmentation can help learn robust representations with a limited scope of augmented samples (Wang et al., 2020a), which we use here.

We report our results in Table 9. The first thing we observe is that the model trained with the augmentation data offered through our approach could preserve a relatively higher performance (FMR 89.13) when testing with the perturbed images generated according to the vanilla model. Since we have shown the perturbed samples have a reasonable transferability in the main manuscript, this result indicates the robustness we brought when training with the perturbed images generated by our approach.

Table 9: Test performances of the model trained in a vanilla manner (denoted as Van.) or with augmentation data offered through our approach (marked by the second column). We report two sets of performances, split by whether the perturbed images are generated according to the vanilla model or the augmented model (marked by the first column).

Data	Model	SA	FMR
Van.	Van.	95.38	57.82
	Aug	87.41	89.13
Aug.	Van.	95.38	58.03
	Aug	87.41	78.61

In addition, when tested with the perturbed images generated according to the augmented model, the augmented model displays a marked resilience (FMR 78.61) in the face of these perturbations compared with the model trained in a vanilla manner (FMR 58.03). Nevertheless, it is noteworthy that the augmented model's performance does exhibit a discernible decline under these circumstances, which once more underscores the efficacy of our approach.

H Grayscale Models

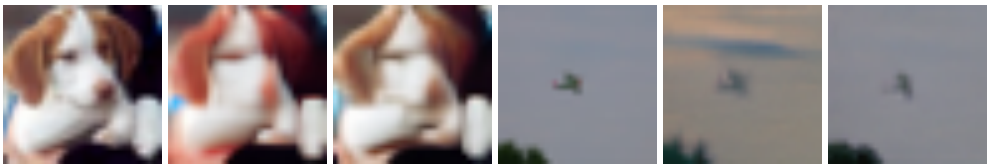
Our previous visualization suggests that a shortcut the perturbed generation system can take is to significantly shift the color of the images, for which a grey-scale model should easily maintain the performance. Thus, we train a grayscale model by changing the ResNet input channel to be 1 and transforming the input images to be grayscale upon feeding into the model. We report the results in Table 10.

Interestingly, we notice that the grayscale model cannot defend against the shift introduced by our system by ignoring the color information. On the contrary, it seems to encourage our system to generate more perturbed images that can lower the performances.

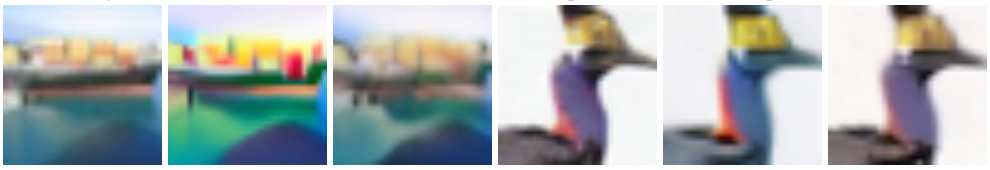
In addition, we visualize some perturbed images generated according to each model and show them in Figure 3. We can see some evidence that the graycale model forces the generation system to focus more on the shape of the object and less of the color of the images. We find it particularly interesting that our system sometimes generates different images differently for different models while the resulting images deceive the respective model to make the same prediction.

Table 10: Test performances of the model trained in a vanilla manner (denoted as Van.) or with grayscale model. We report two sets of performances, split by whether the perturbed images are generated according to the vanilla model or the grayscale one (marked by the first column).

Data	Model	SA	FMR
Van.	Van.	95.38	57.79
	Gray	93.52	66.06
Gray	Van.	95.38	67.48
	Gray	93.52	44.76



(a) dog (b) horse (c) horse (d) plane (e) ship (f) bird



(g) ship (h) car (i) plane (j) bird (k) ship (l) ship

Figure 3: Visualization of the perturbed images generated by our system in evaluating the vanilla model (middle image of each group) and the grayscale model (third image of each group), with the original image shown. The caption for each image is either the original label or the predicted label by the corresponding model.

I Sparse Submodel of VQGAN for Efficient Perturbation

While our method will function properly as described above, we notice that the generation process still has a potential limitation: the bound-free perturbation of VQGAN will sometimes perturb the semantics of the images, generating results that will be rejected by the foundation model later and thus leading to a waste of computational efforts.

To counter this challenge, we use a sparse variable selection method to analyze the embedding dimensions of VQGAN to identify a subset of dimensions that is mainly responsible for the non-semantic variations.

In particular, with a dataset $(\mathbf{X}, \mathbf{Y})$ of n samples, we first use VQGAN to generate a style-transferred dataset $(\mathbf{X}', \mathbf{Y})$. During the generation process, we preserve the latent representations of input samples after the VQGAN encoder in the original dataset. We also preserve the final latent representations before the VQGAN decoder that are quantized after the iterations in the style-transferred dataset. Then, we create a new dataset $(\mathbb{E}, \mathbf{L})$ of $2n$ samples, for each sample $(\mathbf{e}, l) \in (\mathbb{E}, \mathbf{L})$, $\mathbf{e}$ is the latent representation for the sample (from either the original dataset or the style-transferred one), and l is labelled as 0 if the sample is from the original dataset and 1 if the style-transferred dataset.

Then, we train ℓ_1 regularized logistic regression model to classify the samples of $(\mathbb{E}, \mathbf{L})$. With $\mathbf{w}$ denoting the weights of the model, we solve the following problem

$$\arg\min_{\mathbf{w}} \sum_{(\mathbf{e},l)\in(\mathbb{E},\mathbf{L})} l(\mathbf{e}\mathbf{w}, l) + \lambda\|\mathbf{w}\|_1,$$

and the sparse pattern (zeros or not) of $\mathbf{w}$ will inform us about which dimensions are for the style.

We generate the flattened latent representations of input images after the VQGAN Encoder with negative labels. Following Algorithm 1, we generate the flattened final latent representations before the VQGAN decoder with positive labels. Altogether, we form a binary classification dataset where the number of positive and negative samples is balanced. The positive samples are the latent

Table 11: Classification results between vanilla and perturbed images with LASSO.

Data	Sparsity	Test score
MNIST	97.99	78.50
CIFAR-10	98.45	78.00
9-class ImageNet	99.31	72.00
ImageNet	99.32	69.00

representations of perturbed images while the negative samples are the latent representations of input images. We set the split ratio of train and test set to be $0.8 : 0.2$. We perform the explorations on various datasets, i.e. MNIST, CIFAR-10, 9-class ImageNet and ImageNet.

The classification model we consider is LASSO[2] as it enables automatically feature selection with strong interpretability. We set the regularization strength to be 36.36. We adopt saga (Defazio et al., 2014) as the solver to use in the optimization process. The classification results are shown in Table 11.

We observe that the coefficient matrix of features can be far sparser than we expect. We take the result of 9-class ImageNet as an example. Surprisingly, we find that almost 99.31% dimensions in average could be discarded when making judgements. We argue the preserved 0.69% dimensions are highly correlated to VQGAN perturbation. Therefore, we keep the corresponding 99.31% dimensions unchanged and only let the rest 0.69% dimensions participate in computation. Our computation loads could be significantly reduced while still maintain the competitive performance compared with the unmasked version[3].

We conduct the run-time experiments on a single NVIDIA V100 GPU. Following our experiment setting, we evaluate a vanilla ResNet-18 on 9-class ImageNet and a vanilla ResNet-50 on ImageNet. As shown in Table 12, the run-time on ImageNet can be reduced by 28.5% with our sparse VQGAN. Compared with large-scale masked dimensions (*i.e.,* 99.31%), we attribute the relatively incremental run-time improvement (*i.e.,* 12.7% on 9-class ImageNet, 28.5% on ImageNet) to the fact that we have to perform mask and unmask operations each time when calculating the model gradient, which offsets the calculation efficiency brought by the sparse VQGAN to a certain extent.

Table 12: Run-time Comparision between VQGAN and Sparse VQGAN.

Method	Time	
	9-class ImageNet	ImageNet
VQGAN	521.5 ± 1.2s	52602.4 ± 2.7s
Sparse VQGAN	455.4 ± 1.2s	40946.1 ± 2.7s
Improv.	12.7%	28.5%

J Analysis of Samples that are Misclassified by the Model

We notice that, the CLIP model has been influenced by the imbalance sample distributions across the Internet.

In this experiment, we choose a larger step size so that the foundation model may not be able to maintain the image-label structure at the first perturbation. We report the Validation Rate (VR) which is the percentage of images validated by the foundation model that maintains the image-label structure. (In our official configurations, the step size value is small enough that the VR on each dataset is always 1. Therefore, we omit this value in the main experiments.) We present the results on 9-class ImageNet experiment to show the details for each category.

[2] Although LASSO is originally a regression model, we probabilize the regression values to get the final classification results.

[3] We note that the overlapping degree of the preserved dimensions for each dataset is not high, which means that we need to specify these dimensions when facing new datasets.

Table 13: Details of test on 9-class ImageNet for vanilla ResNet-18 (step size is 0.1, computation budget $\mathbf{B}$ is 50)

Type	SA	VR	FMR
Dog	93.33	95.33	17.98
Cat	96.67	94.00	31.55
Frog	85.33	80.67	20.34
Turtle	84.67	78.67	29.03
Bird	91.33	96.00	28.13
Primate	96.00	48.00	62.21
Fish	94.00	76.67	45.33
Crab	96.00	87.33	19.87
Insect	93.33	78.00	33.88
Total	92.30	81.63	30.28

Table 13 shows that the VR values for most categories are still higher than 80%, some even reach 95%, which means we produce sufficient number of perturbed images. However, we notice that the VR value for *primate* images is quite lower compared with other categories, indicating around 52% perturbed *primate* images are blocked by the orcle.

As shown in Table 13, the FMR value for each category significantly drops compared with the SA value, indicating the weakness of trained models. An interesting finding is that the FMR value for *Primate* images are quite higher than other categories, given the fact that more perturbed *Primate* images are blocked by the foundation model. We attribute it to the limitation of foundation models. As the CLIP model has been influenced by the imbalance sample distributions across the Internet, it could only handle easy perturbed samples well. Therefore, the perturbed images preserved would be those that can be easily classified by the models.

Table 14: Details of test on 9-class ImageNet for vanilla ResNet-18 (step size is 0.001, computation budget $\mathbf{B}$ is 50)

Type	SA	VR	FMR
Dog	93.33	100.00	18.09
Cat	96.67	100.00	28.60
Frog	85.33	100.00	20.72
Turtle	84.67	100.00	24.80
Bird	91.33	100.00	27.68
Primate	96.00	100.00	27.11
Fish	94.00	100.00	25.13
Crab	96.00	100.00	19.15
Insect	93.33	100.00	23.16
Total	92.30	100.00	23.94

In our official configuration, we set a relatively smaller step size to perturb the image and obtain enough more perturbed images. As shown in Table 14, using a smaller step size value and enough computation budget barely affect the overall results. In addition, with smller step size, we manage to perturb the image little by little and can get enough more perturbed images (**VR becomes 100 on every category**, indicating that all the images are perturbed and maintained their image-label structure). Admittedly, the foundation model's bias still exists here, e.g., the *Primate* images (FMR = 28.11) are still easier than *Dog* images (FMR = 18.09). However, considering the huge performance gap between the foundation model and the evaluated models, images that are easy for the foundation model are hard enough for the evaluated models (The FMR of *Dogs* and *Primate* images are closer and smaller compared with those in Table 13), which is sufficiently efficacious for real-world applications. Additionally, the employment of an ensemble of multiple foundation models in our methodology serves to provide a further layer of alleviation for the aforementioned issue.

K Discussions on the Societal Bias of Relying on Large Models

K.1 Potential Negative Impacts of Foundation Models

Although the bias incurred by foundation models is less detrimental than the biases arisen from fixed benchmark datasets, a more detailed discussion on the potential negative impacts is necessary. One potential bias of making vision models behave more like the foundation models is that the vision model may inherit the limitations and assumptions of foundation models' training data and objective function. For example, foundation models' training data may not cover all possible visual concepts or scenarios that are relevant to a given task; foundation models' objective function may not align with the desired outcome or evaluation metric of a given task; foundation models' natural language supervision may introduce ambiguities or inconsistencies that affect the model's performance or interpretation. These limitations and assumptions may affect the generalization and robustness of vision models that rely on foundation models. Moreover, we add recent works that especially investigate the bias of foundation models, and guide the readers to it for further warning, e.g., (Menon et al., 2022) and (Zhou et al., 2022b).

K.2 Societal Bias of Relying on Large Models

Moreover, our method relies on large models, where their societal bias is still unclear, therefore a related discussion would be beneficial.

Large-scale models could leverage the rich knowledge and generalization ability encoded in the training stage. However, one potential societal bias of relying on large models' supervision on preserving the perturbed image could be that it would privilege certain groups or perspectives over others based on social or cultural norms. As the data used to train the pre-trained models may be imbalanced, incomplete, or inaccurate, leading to biased representations of certain groups or concepts, the perturbed images preserved by the pre-trained models may reflect stereotypes, or discrimination against certain groups of people based on their race, gender, age, religion, etc., which may be harmful, offensive, or deceptive to the users. Bridging the gap between the pre-trained model and the evaluated vision models will make the vision models inherit the limitations of pre-trained models, which have adverse consequences for people who are affected by them, such as reinforcing stereotypes, discrimination, or exclusion.

We add recent works that investigate the societal bias of large models, and guide the readers to it for further warning, *e.g.,* (Wang et al., 2022a).

L Experiments on the Zero-shot Adversarial Robustness of CLIP

We conduct the following experiment to compare the adversarial vulnerability between CLIP and robust ViT-like model pre-trained checkpoints of XCiT-L12 (Debenedetti et al., 2022) from the RobustBench Leaderboard (Croce et al., 2020). The results are shown in Table 15. We find that the vanilla CLIP shows a better robustness performance under our quick experiments through FGSM attack. However, if we continue the attack process, we will eventually obtain the adversary that changes the CLIP's classification decision to the targeted class.

Table 15: Comparison of the zero-shot adversarial robustness of CLIP with pretrained robust model. We find that CLIP shows a better robustness performance compared with XCiT-L12. We note that the CLIP's classification decision can be changed to the targeted class as attack continues.

Step	**Target loss**		**p[true=0]**		**p[target=1]**	
	CLIP	**XCiT-L12**	**CLIP**	**XCiT-L12**	**CLIP**	**XCiT-L12**
0	8.621	4.712	0.6749	0.7437	0.0052	0.0728
20	2.715	1.605	0.5083	0.4074	0.0986	0.2009
40	2.316	0.8877	0.4007	0.2562	0.1357	0.3116
60	1.684	0.7420	0.2177	0.1407	0.2144	0.4760
80	1.540	0.6520	0.1813	0.1338	0.3335	0.5210

Fortunately, in production, one can use simpler techniques such as gradient masking to protect CLIP's weights from malicious users, thus, the opportunities of the CLIP being attacked from a white-box manner are quite low. In terms of black-box attacks, CLIP actually shows a strong resilience toward the adversarial samples generated for other models, for which we also have some supporting evidence: In Appendix E, we generate the images with the FGSM attack by the tested model. Table 7 shows that initiating with the FGSM adversarial examples barely affects the FMR, which implies that CLIP succeeds in defending these black-box adversarial images and preserving the hard ones such that the FMR does not change significantly (Otherwise, CLIP will discard heavily perturbed images and preserve easy ones with minor perturbation, leading to high FMR values). Furthermore, our approach incorporates an ensemble of foundation models, including robust models such as ConvNext-T-CvSt from the RobustBench Leaderboard, and employs a majority vote mechanism to validate the fidelity of the image-label relationships.

Thus, CLIP, especially when equipped with techniques to protect its weights and gradients, and coupled with an ensemble of robust foundation models, might be the closest one to serve as the ideal foundation models to maintain the image-label structure at this moment.

M List of Evaluated Models

The following lists contains all models we evaluated on various datasets with references and links to the corresponding source code.

M.1 Pretrained VQGAN Model

We use the checkpoint of vqgan_imagenet_f16_16384 from https://heibox.uni-heidelberg.de/d/a7530b09fed84f80a887/

M.2 Pretrained Foundation Models

1. Model weights of ViT-B/32 and usage code are taken from https://github.com/openai/CLIP
2. CoCa (Yu et al., 2022) https://github.com/lucidrains/CoCa-pytorch
3. ConvNeXt-T-CvSt (Singh et al., 2023) https://github.com/nmndeep/revisiting-at

M.3 Timm Models Trained on ImageNet (Wightman, 2019)

Weights are taken from https://github.com/rwightman/pytorch-image-models/tree/master/timm/models

1. ResNet50 (He et al., 2016a)
2. ViT (Dosovitskiy et al., 2020)
3. DeiT (Touvron et al., 2021)
4. Twins (Chu et al., 2021)
5. Visformer (Chen et al., 2021)
6. Swin (Liu et al., 2021)
7. ConvNeXt (Liu et al., 2022)

M.4 Robust ResNet50 Models

1. ResNet50 SIN+IN (Geirhos et al., 2019) https://github.com/rgeirhos/texture-vs-shape
2. ResNet50 ANT (Rusak et al.) https://github.com/bethgelab/game-of-noise
3. ResNet50 ANT+SIN (Rusak et al.) https://github.com/bethgelab/game-of-noise

4. ResNet50 Augmix (Hendrycks et al., 2019) https://github.com/google-research/augmix
5. ResNet50 DeepAugment (Hendrycks et al., 2021a) https://github.com/hendrycks/imagenet-r
6. ResNet50 DeepAugment+Augmix (Hendrycks et al., 2021a) https://github.com/hendrycks/imagenet-r
7. ResNet50 Discrete Adversarial Training (DAT) (Mao et al., 2022b) https://github.com/alibaba/easyrobust

M.5 Additional Image Generators

1. Efficient-VDVAE (Hazami et al., 2022) https://github.com/Rayhane-mamah/Efficient-VDVAE
2. Improved DDPM (Nichol & Dhariwal, 2021) https://github.com/open-mmlab/mmgeneration/tree/master/configs/improved_ddpm
3. ADM (Dhariwal & Nichol, 2021) https://github.com/openai/guided-diffusion
4. StyleGAN (Sauer et al., 2022) https://github.com/autonomousvision/stylegan_xl

M.6 Pretrained XCiT-L12 Model

Model weights of XCiT-L12 (Debenedetti et al., 2022) are taken from https://github.com/dedeswim/vits-robustness-torch

N Leaderboards for Robust Image Model

We launch leaderboards for robust image models. The goal of these leaderboards are as follows:

- To keep on track of state-of-the-art on each adversarial vision task and new model architectures with our dynamic evaluation process.
- To see the comparison of robust vision models at a glance (*e.g.,* performance, speed, size, *etc.*).
- To access their research papers and implementations on different frameworks.

We offer a sample of the robust ImageNet classification leaderboard in supplementary materials.

O Additional Perturbed Image Samples

In Figure 4, we provide additional perturbed images generated according to each model. We have similar observations to Section 4.3. First, the generated perturbed images exhibit diversity that many other superficial factors of the data would be covered, *i.e.,* texture, shape and styles. Second, our method could recognize the model properties, and automatically generate those hard perturbed images to complete the evaluation.

In addition, the generated images show a reasonable transferability in Table 6, indicating tha our method can be potentially used in a broader scope: we can also leverage the method to generate a static set of images and set a benchmark dataset to help the development of robustness methods. Therefore, we also offer two static benchmarks in supplementary materials that are generated based on CNN architecture, *i.e.,* ConvNext and transformer variant, *i.e.,* ViT, respectively.

P Discussion on the realism of the generated images

We notice that some generated images look unnatural, as the generated images being realistic is not part of the optimization function. We acknowledge that making the generated images appear

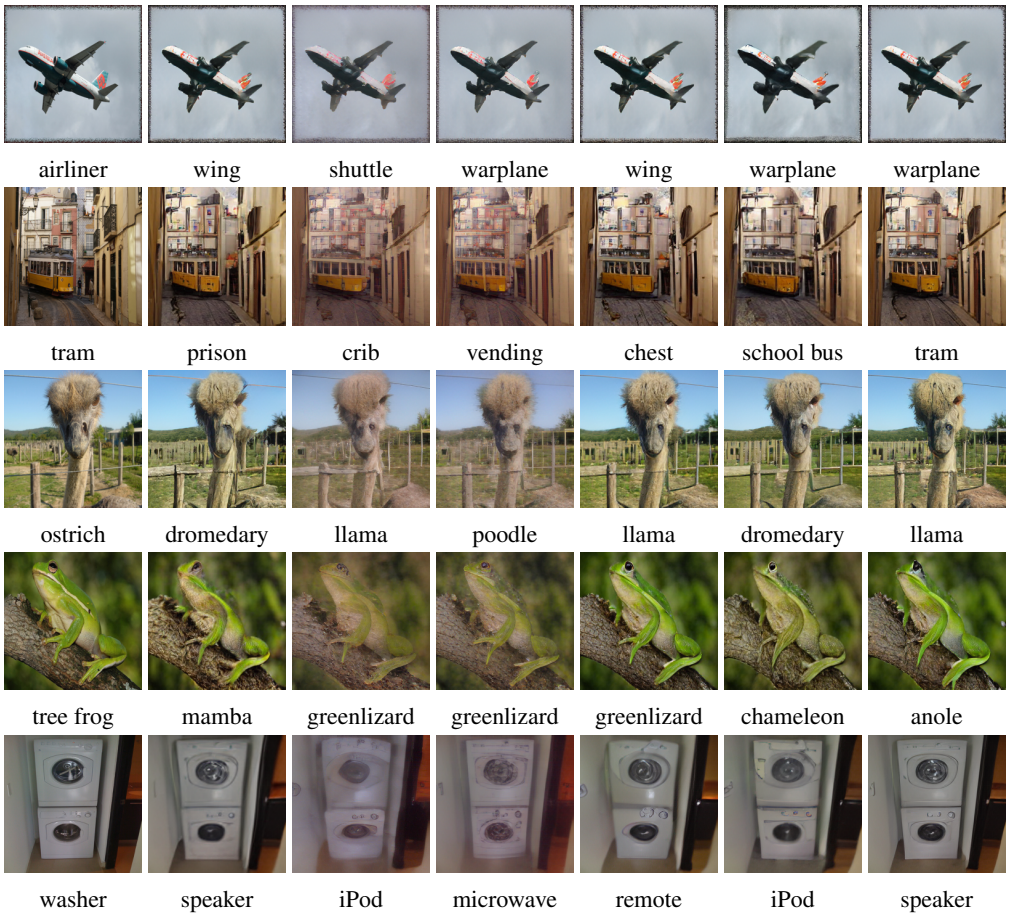


Figure 4: Visualization of the images generated by our system in evaluating the common corruption robust model, with the original image shown (left image of each row). The caption for each image is either the original label or the predicted label by the corresponding model. The evaluated models are SIN, ANT, ANT+SIN, Augmix, DeepAug and DeepAug+AM from left to right.

more natural will be a further desideratum, as this contributes to enhancing the human-perceptible interpretability.

Nonetheless, the current research agenda of the robustness evaluation community is still to encourage the evaluation to expose the model's weakness, such as to expose and eliminate the model's learning of spurious correlation in rare cases.

Similar evidence can be found in (Xiao et al., 2023), where the authors utilize masked images as counterfactual samples for robust fine-tuning. In this paper, the authors argue that masked images can break the spurious correlation between features and labels that may degrade OOD robustness, and that feature-based distillation with the pre-trained model on these counterfactual samples can achieve a better trade-off between IID and OOD performance. According to our second desideratum, our generated counterfactual images might also look unnatural. However, although it appears unnatural, it is beneficial in uncovering and eliminating spurious correlations for enhancing the model robustness.

REFERENCES

Yutong Bai, Jieru Mei, Alan L Yuille, and Cihang Xie. Are transformers more robust than cnns? *Advances in Neural Information Processing Systems*, 34:26831–26843, 2021.

Pedro Ballester and Ricardo Matsumura Araujo. On the performance of googlenet and alexnet applied to sketches. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

Elias Bareinboim, Juan D Correa, Duligur Ibeling, and Thomas Icard. On pearl's hierarchy and the foundations of causal inference. *ACM Special Volume in Honor of Judea Pearl (provisional title)*, 2(3):4, 2020.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In *Advances in neural information processing systems*, pp. 137–144, 2007.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1-2):151–175, 2010.

Digbalay Bose, Rajat Hebbar, Krishna Somandepalli, Haoyang Zhang, Yin Cui, Kree Cole-McLaughlin, Huisheng Wang, and Shrikanth Narayanan. Movieclip: Visual scene recognition in movies. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2083–2092, 2023.

Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, Aleksander Madry, and Alexey Kurakin. On evaluating adversarial robustness. *arXiv preprint arXiv:1902.06705*, 2019.

Zhengsu Chen, Lingxi Xie, Jianwei Niu, Xuefeng Liu, Longhui Wei, and Qi Tian. Visformer: The vision-friendly transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 589–598, 2021.

Xiangxiang Chu, Zhi Tian, Yuqing Wang, Bo Zhang, Haibing Ren, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Twins: Revisiting the design of spatial attention in vision transformers. *Advances in Neural Information Processing Systems*, 34, 2021.

Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. Robustbench: a standardized adversarial robustness benchmark. *arXiv preprint arXiv:2010.09670*, 2020.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, et al. Underspecification presents challenges for credibility in modern machine learning. *arXiv preprint arXiv:2011.03395*, 2020.

Edoardo Debenedetti, Vikash Sehwag, and Prateek Mittal. A light recipe to train robust vision transformers. *arXiv preprint arXiv:2209.07399*, 2022.

Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. *Advances in neural information processing systems*, 27, 2014.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

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.

Richard O Duda, Peter E Hart, and David G Stork. *Pattern classification and scene analysis*, volume 3. Wiley New York, 1973.

Chris Edwards. Malevolent machine learning. *Commun. ACM*, 62(12):13–15, nov 2019. ISSN 0001-0782.

Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Brandon Tran, and Aleksander Madry. Adversarial robustness as a prior for learned representations. *arXiv preprint arXiv:1906.00945*, 2019.

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

Jerome Friedman, Trevor Hastie, Robert Tibshirani, et al. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

Leon Gatys, Alexander S Ecker, and Matthias Bethge. Texture synthesis using convolutional neural networks. *Advances in neural information processing systems*, 28, 2015.

Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Texture and art with deep neural networks. *Current opinion in neurobiology*, 46:178–186, 2017.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples (2014). In *International Conference on Learning Representations*, 2015.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020.

Louay Hazami, Rayhane Mama, and Ragavan Thurairatnam. Efficient-vdvae: Less is more. *arXiv preprint arXiv:2203.13751*, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, 2015.

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

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pp. 630–645. Springer, 2016b.

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

Dan Hendrycks, Norman Mu, Ekin D Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. *arXiv preprint arXiv:1912.02781*, 2019.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8340–8349, 2021a.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15262–15271, 2021b.

Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Pang Wei Koh, Shiori Sagawa, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pp. 5637–5664. PMLR, 2021.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pp. 5542–5550, 2017.

Xihui Liu, Dong Huk Park, Samaneh Azadi, Gong Zhang, Arman Chopikyan, Yuxiao Hu, Humphrey Shi, Anna Rohrbach, and Trevor Darrell. More control for free! image synthesis with semantic diffusion guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 289–299, 2023.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. *arXiv preprint arXiv:2201.03545*, 2022.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pp. 4114–4124. PMLR, 2019.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Divyat Mahajan, Chenhao Tan, and Amit Sharma. Preserving causal constraints in counterfactual explanations for machine learning classifiers. *arXiv preprint arXiv:1912.03277*, 2019.

Chengzhi Mao, Scott Geng, Junfeng Yang, Xin Wang, and Carl Vondrick. Understanding zero-shot adversarial robustness for large-scale models. *arXiv preprint arXiv:2212.07016*, 2022a.

Xiaofeng Mao, Yuefeng Chen, Ranjie Duan, Yao Zhu, Gege Qi, Xiaodan Li, Rong Zhang, Hui Xue, et al. Enhance the visual representation via discrete adversarial training. *Advances in Neural Information Processing Systems*, 35:7520–7533, 2022b.

Diego Marcos, Michele Volpi, and Devis Tuia. Learning rotation invariant convolutional filters for texture classification. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pp. 2012–2017. IEEE, 2016.

Sachit Menon, Ishaan Preetam Chandratreya, and Carl Vondrick. Task bias in vision-language models. *arXiv preprint arXiv:2212.04412*, 2022.

Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, pp. 10–18, 2013.

Nikita Nangia and Samuel R Bowman. Human vs. muppet: A conservative estimate of human performance on the glue benchmark. *arXiv preprint arXiv:1905.10425*, 2019.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pp. 8162–8171. PMLR, 2021.

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

Judea Pearl. *Causality*. Cambridge university press, 2009.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

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

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*, pp. 1278–1286. PMLR, 2014.

Andras Rozsa, Manuel Günther, and Terrance E Boult. Are accuracy and robustness correlated. In *2016 15th IEEE international conference on machine learning and applications (ICMLA)*, pp. 227–232. IEEE, 2016.

Evgenia Rusak, Lukas Schott, Roland Zimmermann, Julian Bitterwolfb, Oliver Bringmann, Matthias Bethge, and Wieland Brendel. Increasing the robustness of dnns against im-age corruptions by playing the game of noise.

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

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

Shibani Santurkar, Dimitris Tsipras, Brandon Tran, Andrew Ilyas, Logan Engstrom, and Aleksander Madry. Image synthesis with a single (robust) classifier. *arXiv preprint arXiv:1906.09453*, 2019.

Axel Sauer, Katja Schwarz, and Andreas Geiger. Stylegan-xl: Scaling stylegan to large diverse datasets. In *ACM SIGGRAPH 2022 Conference Proceedings*, pp. 1–10, 2022.

Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. Improving robustness against common corruptions by covariate shift adaptation. *Advances in Neural Information Processing Systems*, 33:11539–11551, 2020.

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

Naman D Singh, Francesco Croce, and Matthias Hein. Revisiting adversarial training for imagenet: Architectures, training and generalization across threat models. *arXiv preprint arXiv:2303.01870*, 2023.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pp. 2256–2265. PMLR, 2015.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

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

Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 33:18583–18599, 2020.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pp. 10347–10357. PMLR, 2021.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018.

Haohan Wang, Songwei Ge, Eric P Xing, and Zachary C Lipton. Learning robust global representations by penalizing local predictive power. *arXiv preprint arXiv:1905.13549*, 2019.

Haohan Wang, Zeyi Huang, Xindi Wu, and Eric P Xing. Squared ℓ_2 norm as consistency loss for leveraging augmented data to learn robust and invariant representations. *arXiv preprint arXiv:2011.13052*, 2020a.

Haohan Wang, Xindi Wu, Zeyi Huang, and Eric P Xing. High-frequency component helps explain the generalization of convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8684–8694, 2020b.

Haohan Wang, Zeyi Huang, Hanlin Zhang, and Eric Xing. Toward learning human-aligned cross-domain robust models by countering misaligned features. *arXiv preprint arXiv:2111.03740*, 2021.

Junyang Wang, Yi Zhang, and Jitao Sang. Fairclip: Social bias elimination based on attribute prototype learning and representation neutralization. *arXiv preprint arXiv:2210.14562*, 2022a.

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

Ross Wightman. Pytorch image models. `https://github.com/rwightman/pytorch-image-models`, 2019.

Shirley Wu, Mert Yuksekgonul, Linjun Zhang, and James Zou. Discover and cure: Concept-aware mitigation of spurious correlation. *arXiv preprint arXiv:2305.00650*, 2023.

Kevin Xia, Kai-Zhan Lee, Yoshua Bengio, and Elias Bareinboim. The causal-neural connection: Expressiveness, learnability, and inference. *Advances in Neural Information Processing Systems*, 34, 2021.

Yao Xiao, Ziyi Tang, Pengxu Wei, Cong Liu, and Liang Lin. Masked images are counterfactual samples for robust fine-tuning. *arXiv preprint arXiv:2303.03052*, 2023.

Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10687–10698, 2020.

Nanyang Ye, Kaican Li, Lanqing Hong, Haoyue Bai, Yiting Chen, Fengwei Zhou, and Zhenguo Li. Ood-bench: Benchmarking and understanding out-of-distribution generalization datasets and algorithms. *arXiv preprint arXiv:2106.03721*, 2021.

Fisher Yu, Dequan Wang, Evan Shelhamer, and Trevor Darrell. Deep layer aggregation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2403–2412, 2018.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. 2022.

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6023–6032, 2019.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

Qiming Zhang, Yufei Xu, Jing Zhang, and Dacheng Tao. Vitaev2: Vision transformer advanced by exploring inductive bias for image recognition and beyond. *International Journal of Computer Vision*, pp. 1–22, 2023.

Richard Zhang. Making convolutional networks shift-invariant again. In *International conference on machine learning*, pp. 7324–7334. PMLR, 2019.

Tianyuan Zhang and Zhanxing Zhu. Interpreting adversarially trained convolutional neural networks. In *International Conference on Machine Learning*, pp. 7502–7511. PMLR, 2019.

Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 13001–13008, 2020.

Daquan Zhou, Zhiding Yu, Enze Xie, Chaowei Xiao, Animashree Anandkumar, Jiashi Feng, and Jose M Alvarez. Understanding the robustness in vision transformers. In *International Conference on Machine Learning*, pp. 27378–27394. PMLR, 2022a.

Kankan Zhou, Yibin LAI, and Jing Jiang. Vlstereoset: A study of stereotypical bias in pre-trained vision-language models. Association for Computational Linguistics, 2022b.