WITH GREAT BACKBONES COMES GREAT ADVERSARIAL TRANSFERABILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Advancements in self-supervised learning (SSL) for machine vision have enhanced representation robustness and model performance, leading to the emergence of publicly shared pre-trained backbones, such as *ResNet* and *ViT* models tuned with SSL methods like *SimCLR*. Due to the computational and data demands of pre-training, the utilisation of such backbones becomes a strenuous necessity. However, employing backbones may imply adhering to the existing vulnerabilities towards adversarial attacks. Prior research on adversarial robustness typically examines attacks with either full (*white-box*) or no direct access (*black-box*) to the target model, but the adversarial robustness of models tuned on known pre-trained backbones remains largely unexplored. Furthermore, it is unclear which tuning configuration is critical for mitigating exploitation risks. In this work, we systematically study the adversarial robustness of models that use such backbones, evaluating $20,000$ combinations of tuning configurations, including fine-tuning techniques, backbone families, datasets, and attack types. To uncover and exploit vulnerabilities, we propose to use proxy models to transfer adversarial attacks, fine-tuning them with various configurations to simulate different levels of knowledge about the target. Our findings show that proxy-based attacks can outperform strong query-based *black-box* methods with sizeable budgets approaching the effectiveness of *white-box* methods. Critically, we construct a naive "backbone attack", leveraging only the shared backbone, and show that even it can achieve efficacy consistently surpassing *black-box* and closing in towards *white-box* attacks, thus exposing critical risks in model-sharing practices. Finally, our ablations reveal how tuning configuration knowledge impacts attack transferability.

1 INTRODUCTION

Machine vision models pre-trained with massive amounts of data, which utilise self-supervised tuning techniques (Newell & Deng, 2020) are shown to be robust and highly performing (Goyal et al., 2021a; Goldblum et al., 2024) feature-extracting backbones (Elharrouss et al., 2022; Han et al., 2022), which are further used in a variety of tasks, from classification (Atito et al., 2021; Chen et al., 2020b) to semantic segmentation (Ziegler & Asano, 2022). However, creating such backbones incurs substantial data annotation (Jing & Tian, 2020) and computational costs (Han et al., 2022), consequently rendering the use of such publicly available pre-trained backbones the most common and efficient solution for researchers and engineers alike. Prior research has focused on analysing safety and adversarial robustness in different settings w.r.t. knowledge of the target model weights, fine-tuning data, fine-tuning techniques and other tuning configurations – complete knowledge, i.e. *white-box* (Porkodi et al., 2018) vs. no knowledge, i.e. *black-box* (Bhambri et al., 2019).

Although in practice, an attacker can access partial knowledge (Lord et al., 2022; Zhu et al., 2022; Carlini et al., 2022) of how the targeted model was produced, i.e. original backbone weights, tuning recipe, etc., the adversarial robustness of models tuned on a downstream task from a given pre-trained backbone remains largely underexplored. We refer to settings with partial knowledge of the target model tuning configuration as *grey-box* (S. et al., 2018). These types of configurations are important both for research and production settings because with an increased usage (Goldblum et al., 2023) of publicly available pre-trained backbones for downstream applications, we are still incapable of assessing the potential exploitation susceptibility and inherent risks within models tuned on top of them and subsequently enhance future pre-trained backbone sharing practices.

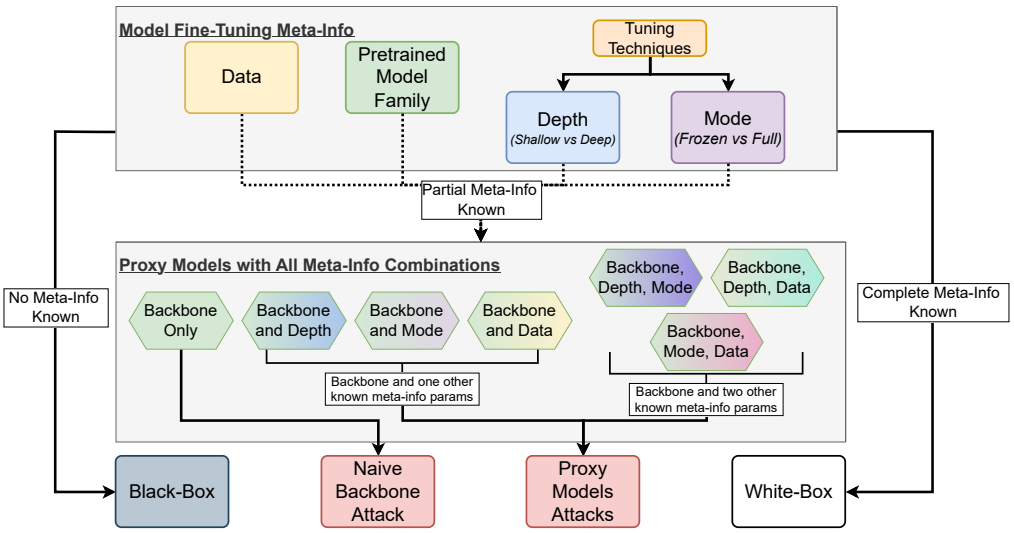

Figure 1: The figure depicts all of the settings used to evaluate adversarial vulnerabilities given different information of the target model construction. From left to right, we simulate exhaustive varying combinations of tuning configurations available from the target model during adversarial attack construction. All of the created proxy models are used separately to assess adversarial transferability.

To address this gap, in our work, we systematically explore the safety from adversarial attacks in models tuned on downstream classification tasks from known publicly available backbones pre-trained with self-supervised objectives. We further explicitly measure the effect of the target model construction configuration by simulating different levels of its availability during the adversarial attack. For this purpose, we initially train 352 diverse models from 21 families of commonly used pre-trained backbones using 4 different fine-tuning techniques and 4 datasets. We fix each of these networks as potential target models and transfer adversarial attacks using all other models produced from the same backbones as proxy surrogates (Qin et al., 2023; Lord et al., 2022) for the construction of adversarial attacks. Each surrogate model simulates varying levels of knowledge availability w.r.t. target model construction configuration on top of the available backbone during adversarial attack construction. This constitutes approximately $20,000$ adversarial transferability comparisons between target and proxy pairs across all model families and configuration variations. By assessing the adversarial transferability of attacks from these surrogate models, we are able to explicitly measure the impact of the availability of each combination of tuning configurations on the final target model during adversarial sample generation, as depicted in Figure 1.

We further explore a naive exploitation method referred to as *backbone attack* that only utilises the pre-trained feature extractor for adversarial sample construction, in this setting. The attack uses projected gradient descent over the representation space to disentangle the features of similar examples. Our results show that both proxy models and even simple *backbone attacks* are capable of surpassing strong query-based *black-box* methods and achieving comparable efficacy to *white-box* performance. The findings indicate that *backbone attacks*, where the attacker lacks knowledge of tuning configuration about the target model, are generally more effective than attempts to generate adversarial samples with limited knowledge. This highlights the vulnerability of models built on publicly available backbones.

Our ablations show that *having access to the weights of the pre-trained backbone is functionally equivalent to possessing all other tuning configurations about the target model when performing adversarial attacks*. We compare these two scenarios and show that both lead to similar vulnerabilities, highlighting the interchangeable nature of these knowledge types in attack effectiveness. Our results emphasise the risks in sharing and deploying pre-trained backbones, particularly concerning the disclosure of configurations. Our experimental framework can be seen in Figure 1.

In summary, our contributions are as follows: **(i)** we formalize and systematically study **grey-box** adversarial attacks, which reflects realistic scenarios where attackers have partial knowledge of target model tuning configuration, such as access to pre-trained backbone weights and/or fine-tuning configuration; **(ii)** we simulate over $20,000$ comparisons of adversarial transferability, evaluating the impact of varying levels of tuning configuration availability about target models during the construction of attacks; **(iii)** we explore a naive attack method, *backbone attacks*, which leverages the pre-trained backbone's representation space for adversarial sample generation, demonstrating that even such a simple approach can achieve stronger performance compared to query-based black-box methods and often approaching white-box attack effectiveness; **(iv)** we show that access to pre-trained backbone weights alone enables adversarial attacks as effectively as access to the full tuning configuration about the target model, emphasizing the inherent vulnerabilities in publicly available pre-trained backbones.

2 Related Work

Self Supervised Learning With the emergence of massive unannotated datasets for machine vision such as YFCC100M (Thomee et al., 2016), ImageNet (Deng et al., 2009), CIFAR (Krizhevsky, 2009) and others, Self Supervised Learning (SSL) techniques (Jing & Tian, 2021) have become increasingly more popular for pre-training vision models (Newell & Deng, 2020). This prompted the creation of various families of SSL objectives, such as colorization prediction (Zhang et al., 2016), jigsaw puzzle solving (Noroozi & Favaro, 2016) with further invariance constraints (Misra & van der Maaten, 2020, PIRL), non-parametric instance discrimination (Wu et al., 2018, NPID, NPID++), unsupervised clustering (Caron et al., 2018), rotation prediction (Gidaris et al., 2018, RotNet), sample clustering with cluster assignment constraints (Caron et al., 2020, SwAV), contrastive representation entanglement (Chen et al., 2020a, SimCLR), self-distillation without labels (Caron et al., 2021, DINO) and others (Jing & Tian, 2021). Numerous architectures, like AlexNet (Krizhevsky et al., 2012), variants of ResNet (He et al., 2016) and visual transformers (Dosovitskiy et al., 2021; Touvron et al., 2021; Ali et al., 2021) were trained using these SSL methods and shared for public use, thus forming the set of widely used pre-trained backbones. We obtain all of these models trained with different self-supervised objectives from their original designated studies summarised in VISSL (Goyal et al., 2021b). An exhaustive list of models is shown in Table 3.

Adversarial Attacks The availability of pre-trained backbones allows for testing them for vulnerabilities towards adversarial attacks, which are learnable imperceptible perturbations generated to mislead models into making incorrect predictions (Szegedy et al., 2014; Goodfellow et al., 2015). Several attack strategies have been studied, including single-step fast gradient descent (Goodfellow et al., 2014; Kurakin et al., 2017, FGSM), and computationally more expensive optimization-based attacks, such as projected gradient descent based attacks (Madry et al., 2018, PGD), CW (Carlini & Wagner, 2017), JSMA (Papernot et al., 2017), and others (Dong et al., 2018; Moosavi-Dezfooli et al., 2016; Madry et al., 2018; Ma et al., 2023). All of these attacks assume complete access to the target model, which is known as the *white-box* (Papernot et al., 2017) setting. These attacks can be *targeted* to confuse the model to infer a specific wrong class or *untargeted*, aiming to make them infer any incorrect label. However, an opposite setting with no information, referred to as *black-box* (Papernot et al., 2017), has also been explored as a more common setting during adversarial attack construction. These methods involve attempts at gradient estimation (Chen et al., 2017; Ilyas et al., 2018; Bhagoji et al., 2018), adversarial transferability (Papernot et al., 2017; Chen et al., 2020c), local search (Narodytska & Kasiviswanathan, 2016; Brendel et al., 2018; Li et al., 2019; Moon et al., 2019), combinatorial perturbations (Moon et al., 2019) and others (Bhambri et al., 2019). However, a great portion of these methods also require massive sample query budgets ranging from $\left[10^3, 10^5\right]$ queries, or computational resources for creating each adversarial sample (Bhambri et al., 2019). Compared to these, we introduce a novel setup with the knowledge of the pre-trained backbone and varying levels of partially known target model tuning configuration during adversarial attack construction, which we refer to as *grey-box*. This setup reflects common scenarios where attackers have partial knowledge of the target model tuning configuration, allowing them to systematically assess the effect of this knowledge on adversarial transferability and show the risks in the current model-sharing practices. We show that even simple naive attacks are more capable of exploiting models without the need for a sizable query budget compared to black-box attacks.

Adversarial Transferability Our work is also aligned with adversarial transferability, where adversarial examples generated for one model can mislead other models, even without access to the target model weights or training data. This property poses significant security concerns, as it allows for effective black-box attacks on systems with no direct access (Papernot et al., 2017; Ilyas et al., 2018). Efforts can be divided into *generation-based* and *optimisation* methods. Generative methods have emerged as an alternative to iterative attacks, where adversarial generators are trained to produce transferable perturbations. For instance, Poursaeed et al. (2018) employs autoencoders trained on white-box models to generate adversarial examples. Most attacks aiming at adversarial transferability strongly depend on the availability of data from the target domain (Carlini & Wagner, 2017; Papernot et al., 2017), although attempts at improving the transferability of baseline adversarial samples have also been explored (Li et al., 2020; Zhang et al., 2022; Li et al., 2023; 2024; Naseer et al., 2020a). However, although current adversarial transferability methods claim to produce massive vulnerabilities in machine vision models, Katzir & Elovici (2021) examines the practical implications of adversarial transferability, which are frequently overstated. That study demonstrates that it is nearly impossible to reliably predict whether a specific adversarial example will transfer to an unseen target model in a black-box setting. This perspective shows the importance of systematically evaluating transferability in realistic settings, including scenarios where attackers are sensitive to the cost of failed attempts. In our study, we offer a novel systematic approach to explicitly assess the adversarial transferability with varying levels of configuration knowledge.

3 METHODOLOGY

Preliminaries For consistency, we employ the following notation. We denote each dataset as $\mathcal{D} = \{\mathcal{X}, \mathcal{Y}\}$; where $\mathcal{X} = \{x_1, \ldots, x_{|\mathcal{D}|}\}$ is a set of images, with $x_i \in \mathcal{R}^{H \times W \times C}$; where H,W and C are the height, width and the channels of the image accordingly and $\mathcal{Y} = \{y_1 \ldots y_n\}$ is used as the set of ground truth labels. We denote the training, validation and testing splits per task as $\mathcal{D} = \{\mathcal{D}_{train}, \mathcal{D}_{val}, \mathcal{D}_{test}\}$. A *model* is defined as a tuple $\mathcal{M} = \mathcal{M}(\mathcal{D}, \mathcal{W}, \mathcal{B}, \mathcal{F})$, where $\mathcal{D}$ contains the dataset used for training, $\mathcal{W}$ are the weights of the trained model and $\mathcal{B}$ is the pre-trained back-bone $\mathcal{B}(\mathcal{W}_\mathcal{B})$ with available weights $\mathcal{W}_\mathcal{B}$. The notation $\mathcal{F}(\mathcal{T}, \mathcal{Z})$, where $\mathcal{T}$ encodes the *mode* of tuning (e.g., full fine-tuning, partial fine-tuning, etc.) and $\mathcal{Z}$ the *depth* of tuning of the final classifier on top of the backbone.

Tuning configuration variations We define the variations of the available configuration about the target model $\mathcal{M}$ during an adversarial attack as a *unit of release* $\mathcal{R} = \mathcal{R}(\mathcal{M}(\mathcal{D}, \mathcal{W}, \mathcal{B}(\mathcal{W}_\mathcal{B}), \mathcal{F}(\mathcal{T}, \mathcal{Z})))$. For example, if the target fine-tuning mode $\mathcal{Z}^{\text{target}}$ and dataset $\mathcal{D}^{\text{target}}$ are not known, the unit of release will be $\mathcal{R} = \mathcal{R}(\mathcal{M}(*, \mathcal{W}, \mathcal{B}(\mathcal{W}_\mathcal{B}), \mathcal{F}(\mathcal{T}, *)))$. Note that the *black-box* setting will correspond to the unit of release $\mathcal{R}(\mathcal{M}(*, *, *, *, *))$ and the *white-box* setting to $\mathcal{R}(\mathcal{M}(\mathcal{D}, \mathcal{W}, \mathcal{B}(\mathcal{W}_\mathcal{B}), \mathcal{F}(\mathcal{T}, \mathcal{Z})))$, all the variations between these are considered *grey-box*. When discussing any experiments within the *grey-box* setup, we assume the minimal unit of release contains knowledge about at least the pre-trained backbone i.e. $\mathcal{R}(\mathcal{M}(*, *, \mathcal{B}(\mathcal{W}_\mathcal{B}), *)$.

Adversarial Attacks with Proxy Models To test the adversarial robustness of the models trained from the same pre-trained backbone, we create a set of proxy models $\mathcal{M}^{\text{proxy}} = \{\mathcal{M}_1^{\text{proxy}} \ldots \mathcal{M}_v^{\text{proxy}}\}$ given the pre-trained backbone $\mathcal{B}$, where v is the number of all possible units of release between *black-box* and *white-box* settings that include the backbone. For each proxy model $\mathcal{M}_i^{\text{proxy}}$ with its designated configuration unit of release $\mathcal{R}_i$, we use an adversarial attack $\mathcal{A}$ to generate adversarial noise and further transfer it to the target model $\mathcal{M}^{\text{target}}$. This means that given an example image x with a label y, target and proxy models $\mathcal{M}^{\text{target}}$, $\mathcal{M}^{\text{proxy}}$ we want to produce a sample x' that would fool the target model, such that $\arg\max \mathcal{M}^{\text{target}}(x') \neq y$. If we are using a targeted attack, we want $\mathcal{M}^{\text{target}}(x') = t$ where t is the targeted class different from the ground truth $t \neq c_{gt}$. After creating the adversarial attack for each sample in $\mathcal{D}_{\text{test}}^{\text{proxy}}$ and $\mathcal{D}_{\text{test}}^{\text{target}}$, we evaluate the success rate of the attack and the success rate of the transferability to the target model. To measure the success and robustness of the adversarial attack and its transferability, we define the following metrics:

Attack Success Rate (ASR). The proportion of adversarial examples that fool the proxy model $\mathcal{M}_i^{\text{proxy}}$:

$$\text{ASR}_i = \frac{1}{|\mathcal{D}_{\text{test}}^{\text{proxy}}|} \sum_{x \in \mathcal{D}_{\text{test}}^{\text{proxy}}} \mathbb{I}\left[\mathcal{M}_i^{\text{proxy}}(x') \neq y\right], \tag{1}$$

Algorithm 1 Backbone Attack

Input: Model backbone $\mathcal{B}$, clean image x_0, perturbation bound ϵ, step size α, number of steps T, distance function $\mathcal{L}_{\text{cosine}}$, random start flag
Output: Adversarial image x_{adv}
Initialization:
$x_{\text{adv}} \leftarrow x_0$
if *random start* **then**
  $x_{\text{adv}} \leftarrow x_{\text{adv}} + \text{Uniform}(-\epsilon, \epsilon)$
  $x_{\text{adv}} \leftarrow \text{Clip}(x_{\text{adv}}, 0, 1)$
end
Fixed Original Image Representation:
$z_0 \leftarrow StopGrad(\mathcal{B}(x_0))$

for $t = 1$ ***to*** T **do**
  Forward Pass:
    $z_{\text{adv}} \leftarrow \mathcal{B}(x_{\text{adv}})$ `// Adversarial image representation`
  Compute Loss and Gradient:
    $\mathcal{L} \leftarrow 1 - \cos(z_{\text{adv}}, z_0)$ `// Distance loss`
  $g \leftarrow \nabla_{x_{\text{adv}}} \mathcal{L}$ `// Gradient w.r.t` x_{adv}
  Update Adversarial Image:
    $x_{\text{adv}} \leftarrow x_{\text{adv}} + \alpha \cdot \text{sign}(g)$ `// PGD step`
  Projection:
    $\delta \leftarrow \text{Clip}(x_{\text{adv}} - x_0, -\epsilon, \epsilon)$ `// Project perturbation into` ℓ_∞`-ball with perturbation budget` ϵ
  $x_{\text{adv}} \leftarrow \text{Clip}(x_0 + \delta, 0, 1)$ `// pixel range`
end
return x_{adv}

where $\mathbb{I}[\cdot]$ is the indicator function.

Transfer Success Rate (TSR). The proportion of adversarial examples generated by $\mathcal{M}_i^{\text{proxy}}$ that also fool the target model $\mathcal{M}^{\text{target}}$:

$$\text{TSR}_i = \frac{1}{|\mathcal{D}_{\text{test}}^{\text{target}}|} \sum_{x \in \mathcal{D}_{\text{test}}^{\text{target}}} \mathbb{I}\left[\mathcal{M}^{\text{target}}(x') \neq y\right]. \tag{2}$$

This setup allows us to explicitly quantify how the availability of diverse configuration combinations explicitly impacts the adversarial transferability of the given model, thus highlighting the risks in the model-sharing practices. A visual depiction of this can be seen in Figure 1.

3.1 Backbone Attack

To test the vulnerabilities associated with publicly available pre-trained feature extractors, we construct a *backbone attack*, which only utilises the known backbone $\mathcal{B}$ of the model $\mathcal{M}^{\text{target}}$. The aim, similar to the prior paragraph, is to create an adversarial attack from $\mathcal{B}$ to transfer to the target model $\mathcal{M}^{\text{target}}$. To do this, we use a Projected Gradient Descent-based method (Madry et al., 2018, PGD), where the attack iteratively perturbs the input images in order to maximize the distance between the feature representations of the clean input and the adversarial input, as derived from the backbone $\mathcal{B}$. More formally, let x and $\tilde{x}$ represent the clean input and adversarial input, respectively. The attack iteratively refines $\tilde{x}$ such that:

$$\tilde{x}_{t+1} = \text{Proj}_{\mathcal{S}}\left(\tilde{x}_t + \alpha \cdot \text{sign}\left(\nabla_{\tilde{x}_t} \mathcal{L}_{\mathcal{B}}(x, \tilde{x}_t)\right)\right), \tag{3}$$

where $\mathcal{L}_{\mathcal{B}}$ is the loss function defined to measure the distance between the feature representations of the clean and adversarial inputs. The backbone representations $f_{\mathcal{B}}$ are extracted as $f_{\mathcal{B}}(x) = \mathcal{B}(x)$, and the differentiable loss can be formulated as:

$$\mathcal{L}_{\mathcal{B}}(x, \tilde{x}) = 1 - \cos\left(f_{\mathcal{B}}(x), f_{\mathcal{B}}(\tilde{x})\right), \tag{4}$$

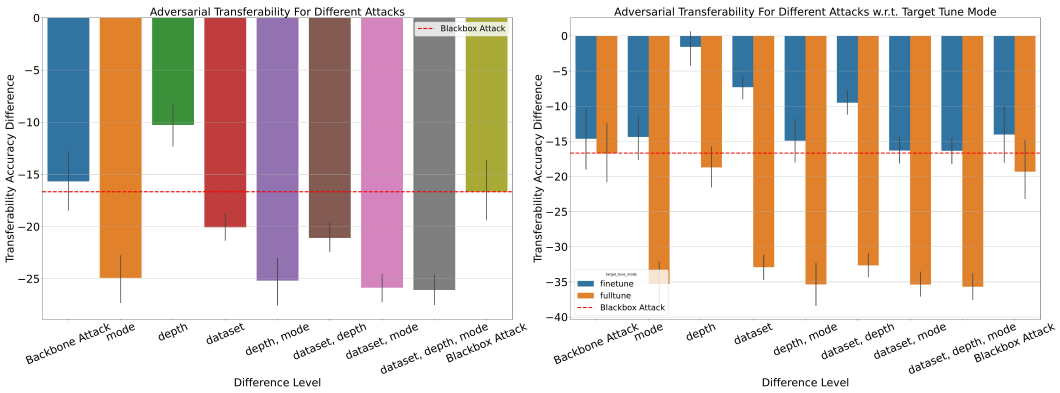

Figure 2: The impact of the **unavailability**, i.e. difference from the target model *white-box* performance, of all tuning configurations on adversarial transferability during proxy attack construction and the backbone attack. The results show the average difference from the *white-box* in transferability using PGD with a higher budget (left) and the segmentation w.r.t. in the target training mode (right).

where $\cos(\cdot,\cdot)$ represents the cosine similarity between the two feature vectors. To prevent gradient computation from propagating to the clean representation $f_{\mathcal{B}}(x)$, we utilize a stop-gradient operation $\tilde{f}_{\mathcal{B}}(x) = \text{SG}(f_{\mathcal{B}}(x))$. The adversarial input $\tilde{x}$ is initialized with a random perturbation within the ℓ_∞ ball of radius ϵ, and the updates are iteratively projected back onto this ball using the $\text{Proj}_{\mathcal{S}}$ operator:

$$\begin{aligned} &\text{Proj}_{\mathcal{S}}(\tilde{x}) = \text{clip}\left(x + \delta, 0, 1\right), \\ &\text{where} \quad \delta = \text{clip}\left(\tilde{x} - x, -\epsilon, \epsilon\right). \end{aligned} \tag{5}$$

The pseudo-code of the complete process can be seen in Algorithm 1. In summary, the backbone attack focuses solely on the backbone $\mathcal{B}$, without requiring any knowledge of the full target model $\mathcal{M}^{\text{target}}$, thereby revealing vulnerabilities inherent to publicly available feature extractors. A form of this algorithm has been utilised as a naive self-supervised perturbation generation component in adversarial defence training (Naseer et al., 2020b, NPR), however, it has not been explored individually. We only use this attack to showcase that even naive backbone exploitation methods can have significant adversarial transferability.

4 Experimental Setup

Image classification datasets Through our study, we use 4 datasets covering both classical and domain-specific classification benchmarks, such as CIFAR-10 and CIFAR-100 (Beyer et al., 2020), Oxford-IIIT Pets (Parkhi et al., 2012) and Oxford Flowers-102 (Nilsback & Zisserman, 2008). We train the proxy and target model variations on each one of the datasets using the recipe and hyperparameters by (Kolesnikov et al., 2020), reproducing the state-of-the-art model performance results (Dosovitskiy et al., 2020; Yu et al., 2022; Bruno et al., 2022; Foret et al., 2020).

	Original Entropy		Adversarial Entropy	
Metadata type	**F-Statistic**	**P-Value**	**F-Statistic**	**P-Value**
Target Tune Mode	0.00	0.96	1238.7	0.0
Proxy Tune Mode	0.02	0.88	0.5	0.4
Target Dataset	2812.25	0.00	1184.1	0.0
Proxy Dataset	8.31	0.00	5.0	0.0
Target Tune Depth	5.64	0.01	0.36	0
Proxy Tune Depth	0.08	0.77	0.00	0

Table 1: Variance analysis of entropy values across categorical variables. The table shows F-statistics and p-values for both original and adversarial entropy means. Significant p-values ($p < 0.05$) show notable variations in entropy across tuning configurations.

Model variations We use 21 different models tuned from 5 architectures, 9 self-supervised objectives and 3 pre-training datasets. A detailed overview of these can be seen in Table 3 in Section A.1.

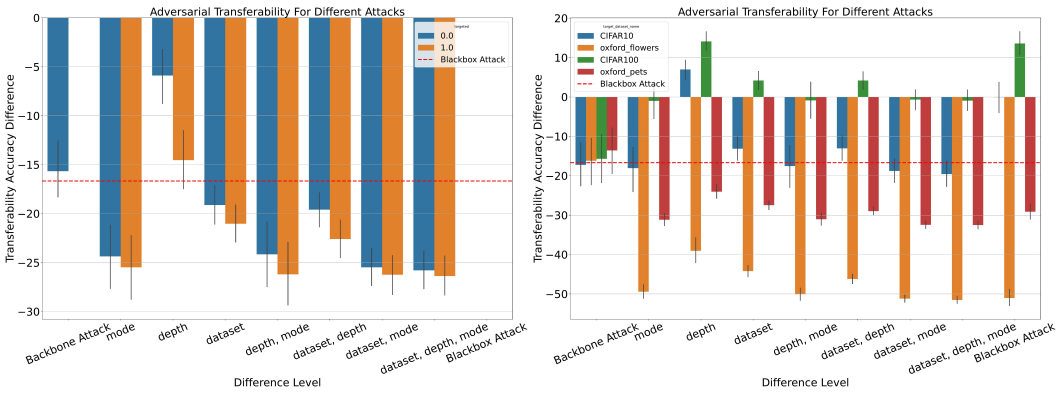

Figure 3: The impact of the **unavailability**, i.e. difference from the target model *white-box* performance, of all tuning configurations on adversarial transferability during proxy attack construction and the backbone attack. The results show the average transferability for PGD with a higher budget for targeted vs untargeted attacks (left) and the segmentation w.r.t. the target training dataset (right).

Model Fine-tuning Variations For training the proxy and target models, we employ two *modes* of training $\mathcal{T}$, with full-tuning of the weights and with fine-tuning only the last added classification layers on top of the pre-trained backbone. We also define the depth of tuning $\mathcal{Z}$ as the number of classification layers added on top of the pre-trained backbone. We use $\{1, 3\}$ final layers, corresponding to *shallow* and *deep* tuning settings.

Adversarial Attacks To assess the success rate of *white-box* adversarial attacks and the adversarial transferability from the proxy models, we employ FGSM (Goodfellow et al., 2015) and PGD (Madry et al., 2018). We use standard attack hyper-parameters introduced in parallel adversarial transferability studies (Waseda et al., 2023; Naseer et al., 2022). For a fair comparison, we also use the same values for our *backbone-attack*. We also impose a standard perturbation budget $\epsilon \leq \frac{8}{255}$ in line with prior studies (Naseer et al., 2022) outline in Algorithm 1. To show that our results are consistent even with a higher computational budget, we report the results of PGD with 4 times more iterations per sample for *white-box*, proxy and *backbone* attack experiments. For *black-box* experiments, we use the Square attack (Andriushchenko et al., 2020), which is a query-efficient method that uses a random search through adversarial sample construction. To standardise the query budget for all architectures and simulate real-world constraints, we allow 10 queries of the target model per sample. The information about the used computational resource can be found in Section A.2.

5 Results

5.1 What configuration matters

To quantify the impact of each possible configuration availability along with the backbone knowledge during adversarial attack construction, we compute the difference between the adversarial attack success rate (ASR) for the target model and the transferability success rate (TSR) from a proxy model, trained from the same backbone, with partial information. We report the results obtained with the PGD attack trained with higher iteration steps per sample as that is more representative for measuring the adversarial attack success in *white-box* and *grey-box* settings.

Which configuration is important? Our results in Figure 2 show that the most significant performance decay compared to a *white-box* attack performance occurs when the attacker is unaware of the *mode* of the training of the target model, i.e. if it is trained with complete parameters or only tunes the last classification layers. The second most impactful knowledge for attack construction is the availability of the target tuning *dataset*. The *depth* of the tuning is the least important knowledge for obtaining a transferable attack. We further show in the right part of Figure 2 that models that fine-tune the last classification layers can be trivially exploited with transferable attacks, achieving results significantly better than strong black-box exploitation and closing white-box attack performance. It

is, however, apparent that training all of the model weights substantially decreases the efficiency of proxy attacks, with almost no correlation towards configuration availability. We further show that our results remain consistent w.r.t. the choice of the dataset, and regardless if the adversarial attack is targeted or untargeted as seen in Figure 3. It is interesting to note that for datasets with more domain-specific content, such as Oxford-IIIT Pets and Oxford Flowers-102, the effectiveness of the proxy attack dwindles, although these datasets are much less diverse compared to CIFAR-100.

Tuning configuration impacts the quality of adversarial attacks We also want to measure the effectiveness of the adversarial attack and the impact of the tuning configuration on it by quantifying how the generated adversarial sample has shifted the decision-making of the model. To do this, we compute the entropy of the final softmax layer for each original sample and its adversarial counterpart, and perform a complete ANOVA variance analysis (St et al., 1989) of the entropy distribution. This analysis, presented in Table 1, tests whether the means of entropies from the original and adversarial images differ significantly across the groups of available tuning configurations. A perfect attack would produce a sample that does not majorly impact the entropy of the model. The analysis reveals that the target dataset, and tuning mode significantly influence entropy, particularly in adversarial scenarios. This finding suggests that while this configuration aids in crafting effective adversarial samples, it also plays a critical role in amplifying entropy shifts, thereby making these adversarial samples more detectable.

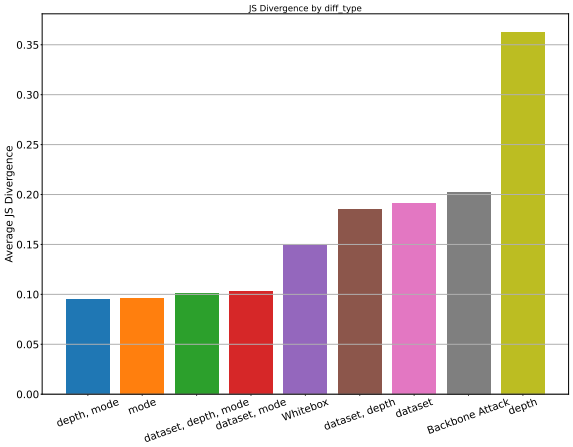

Figure 4: Impact of the **unavailability** of each tuning configuration on model decision-making. Higher JS divergence implies a bigger change in final classification.

To quantify the impact of the availability of tuning configuration during the construction of attacks on the decision-making of the model, we also compute the Jensen-Shannon Divergence (Menéndez et al., 1997) between the output softmax distributions of the model produced for original samples and their adversarial counterparts, seen in Figure 4. High JS divergence suggests a strong attack, as the adversarial example causes a significant shift in the model's predicted probabilities, with minimal changes to the input sample under an imposed perturbation budget ϵ. Our results show that not knowing the *mode* of the target model training causes the most degradation in constructing successful adversarial samples with proxy attacks. The second most important fact is the choice of the target *dataset*, while the *depth* of the final classification layers does not seem to be impactful for creating adversarial samples. Figure 4 reveals a critical insight: proxy attacks, even when constructed without knowledge of the target model's *dataset* or *depth*, can generate adversarial samples that induce more pronounced distribution shifts than *white-box* attacks. In other words, attackers do need to have access to the training dataset or model classification depth to craft adversarial samples capable of significantly disrupting the target model's decision-making process.

5.2 Backbone-attacks

To test the extent of the vulnerabilities that the knowledge of the pre-trained backbone can cause, we evaluate a naive exploitation method, *backbone attack*, which only uses the pre-trained feature extractor for adversarial sample construction. Our results in Figure 2 and Figure 3 show that *backbone attacks* are highly effective at producing transferable adversarial samples regardless of the target model tuning *mode*, *dataset* or classification layer *depth*. This naive attack shows significantly higher transferability compared to a strong *black-box* attack with a sizeable query and iteration budget and almost all *proxy attacks*. The results are consistent across all configuration variations, showing that even a naive attack can exploit the target model vulnerabilities closely to a *white-box* setting, given the knowledge of the pre-trained backbone. Moreover, in Figure 4, we see that the adversarial samples produced from this attack, on average, cause a bigger shift in the model's decision-making

compared to *white-box attacks*. This indicates that backbone attacks amplify the uncertainty in the target model's predictions, making them more disruptive than conventional *white-box* attacks. A concerning aspect of backbone attacks is their effectiveness in resource-constrained environments. Unlike black-box attacks, which often require extensive computation or iterative querying, backbone attacks can be executed with minimal resources, leveraging pre-trained models freely available in public repositories. This ease of implementation raises concerns, as it lowers the barrier for malicious actors to exploit adversarial vulnerabilities.

5.3 KNOWING THE WEIGHTS VS KNOWING EVERYTHING ELSE

To isolate the impact of pre-trained backbone knowledge in adversarial transferability, we take two ResNet-50 SwAV backbones pre-trained with different batch sizes and further tuned with identical configuration variations. This allows for the production of two sets of models with matching training configurations but varying weights; one set is chosen as the target, and the other as the proxy model. We aim to compare the adversarial transferability of the attacks from the set of proxies towards their matching targets with the backbone attacks. This allows us to simulate conditions where adversaries either know all configurations but lack the weights or have access to the backbone weights alone.

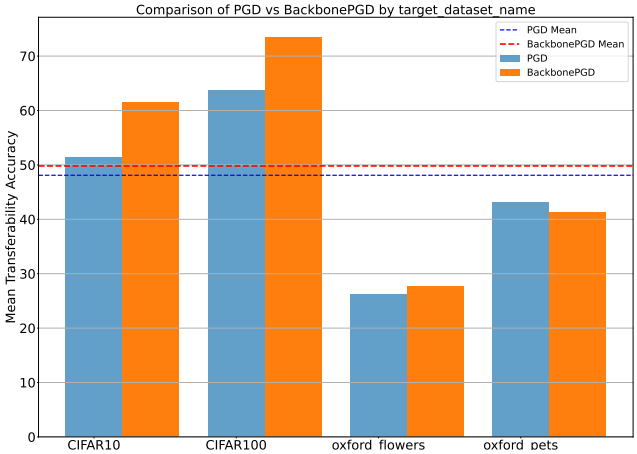

Figure 5: Scenarios where adversaries either lack backbone weights or only possess them. The latter is shown as *BackbonePGD* (SwaV ResNet-50).

Our results in Figure 5 show that the knowledge of the pre-trained backbone is, on average, a stronger or at least an equivalent signal for producing adversarially transferable attacks compared to possessing all of the training configurations without the knowledge of the weights. The results are consistent across all datasets, with domain-specific datasets showing marginal differences in adversarial transferability between the two scenarios. This means that possessing information about only the target model backbone is equivalent to knowing all of the training configurations for constructing transferable adversarial samples.

6 CONCLUSIONS

We investigated the vulnerabilities of machine vision models fine-tuned from publicly available pre-trained backbones under a formalised *grey-box* adversarial setting. We systematically measured the effect of varying levels of training configuration availability for constructing transferable adversarial attacks. We also explored a naive *backbone attack* method in this setting, showing that access to backbone weights is sufficient for obtaining adversarial attacks significantly better than query-based *black-box* settings and comparable to white-box performance. We found that these attacks often induce more drastic shifts in the model's decision-making compared to white-box attacks. We demonstrated that access to backbone weights is equivalent in effectiveness to possessing all tuning configurations about the target model, making public backbones a critical security concern. Our results highlight the risks associated with sharing pre-trained backbones, as they enable attackers to craft highly effective adversarial samples, even with minimal additional information. These findings underscore the need for more thought-out practices in sharing pre-trained backbones to mitigate the inherent vulnerabilities exposed by adversarial transferability.

Ethics Statement

We confirm that our experiments respect privacy, avoid misuse, disclose limitations and potential harms, and acknowledge societal impacts. All datasets used were obtained under proper licenses or permissions, and any use of adversarial methods is justified and documented to alert downstream users of risks.

Reproducibility Report

To reproduce the results of our study, we provide the complete codebase, processing pipelines and hyperparameters for each dataset. We also make the rigorous details and checkpoints of all of the models in our study across all of the datasets publicly available for further experimentation and exploration.

References

Alaaeldin Ali, Hugo Touvron, Mathilde Caron, Piotr Bojanowski, Matthijs Douze, Armand Joulin, Ivan Laptev, Natalia Neverova, Gabriel Synnaeve, Jakob Verbeek, and Hervé Jégou. Xcit: Cross-covariance image transformers. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 20014–20027, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/a655fbe4b8d7439994aa37ddad80de56-Abstract.html.

Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: A query-efficient black-box adversarial attack via random search. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (eds.), *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXIII*, volume 12368 of *Lecture Notes in Computer Science*, pp. 484–501. Springer, 2020. doi: 10.1007/978-3-030-58592-1_29. URL https://doi.org/10.1007/978-3-030-58592-1_29.

Sara Atito, Muhammad Awais, and Josef Kittler. Sit: Self-supervised vision transformer. *arXiv preprint arXiv:2104.03602*, 2021.

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.

Arjun Nitin Bhagoji, Warren He, Bo Li, and Dawn Song. Practical black-box attacks on deep neural networks using efficient query mechanisms. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (eds.), *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XII*, volume 11216 of *Lecture Notes in Computer Science*, pp. 158–174. Springer, 2018. doi: 10.1007/978-3-030-01258-8_10. URL https://doi.org/10.1007/978-3-030-01258-8_10.

Siddhant Bhambri, Sumanyu Muku, Avinash Tulasi, and Arun Balaji Buduru. A survey of black-box adversarial attacks on computer vision models. *arXiv preprint arXiv:1912.01667*, 2019.

Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=SyZI0GWCZ.

Antonio Bruno, Davide Moroni, and Massimo Martinelli. Efficient adaptive ensembling for image classification. *arXiv preprint arXiv:2206.07394*, 2022.

Nicholas Carlini and David A. Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In Bhavani Thuraisingham, Battista Biggio, David Mandell Freeman, Brad Miller, and Arunesh Sinha (eds.), *Proceedings of the 10th ACM Workshop on Artificial Intelligence*

and Security, AISec@CCS 2017, Dallas, TX, USA, November 3, 2017, pp. 3–14. ACM, 2017. doi: 10.1145/3128572.3140444. URL https://doi.org/10.1145/3128572.3140444.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramèr. Membership inference attacks from first principles. In *43rd IEEE Symposium on Security and Privacy, SP 2022, San Francisco, CA, USA, May 22-26, 2022*, pp. 1897–1914. IEEE, 2022. doi: 10.1109/SP46214.2022.9833649. URL https://doi.org/10.1109/SP46214.2022.9833649.

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (eds.), *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIV*, volume 11218 of *Lecture Notes in Computer Science*, pp. 139–156. Springer, 2018. doi: 10.1007/978-3-030-01264-9_9. URL https://doi.org/10.1007/978-3-030-01264-9_9.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/70feb62b69f16e0238f741fab228fec2-Abstract.html.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 9630–9640. IEEE, 2021. doi: 10.1109/ICCV48922.2021.00951. URL https://doi.org/10.1109/ICCV48922.2021.00951.

Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. ZOO: zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In Bhavani Thuraisingham, Battista Biggio, David Mandell Freeman, Brad Miller, and Arunesh Sinha (eds.), *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, AISec@CCS 2017, Dallas, TX, USA, November 3, 2017*, pp. 15–26. ACM, 2017. doi: 10.1145/3128572.3140448. URL https://doi.org/10.1145/3128572.3140448.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR, 2020a. URL http://proceedings.mlr.press/v119/chen20j.html.

Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33:22243–22255, 2020b.

Weilun Chen, Zhaoxiang Zhang, Xiaolin Hu, and Baoyuan Wu. Boosting decision-based black-box adversarial attacks with random sign flip. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (eds.), *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XV*, volume 12360 of *Lecture Notes in Computer Science*, pp. 276–293. Springer, 2020c. doi: 10.1007/978-3-030-58555-6_17. URL https://doi.org/10.1007/978-3-030-58555-6_17.

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.

Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 9185–9193. Computer Vision Foundation / IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00957. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Dong_Boosting_Adversarial_Attacks_CVPR_2018_paper.html.

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.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Omar Elharrouss, Younes Akbari, Noor Almaadeed, and Somaya Al-Maadeed. Backbones-review: Feature extraction networks for deep learning and deep reinforcement learning approaches. *arXiv preprint arXiv:2206.08016*, 2022.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*, 2020.

Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=S1v4N2l0-.

Micah Goldblum, Hossein Souri, Renkun Ni, Manli Shu, Viraj Prabhu, Gowthami Somepalli, Prithvijit Chattopadhyay, Mark Ibrahim, Adrien Bardes, Judy Hoffman, Rama Chellappa, Andrew Gordon Wilson, and Tom Goldstein. Battle of the backbones: A large-scale comparison of pretrained models across computer vision tasks. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/5d9571470bb750f0e2325a030016f63f-Abstract-Datasets_and_Benchmarks.html.

Micah Goldblum, Hossein Souri, Renkun Ni, Manli Shu, Viraj Prabhu, Gowthami Somepalli, Prithvijit Chattopadhyay, Mark Ibrahim, Adrien Bardes, Judy Hoffman, et al. Battle of the backbones: A large-scale comparison of pretrained models across computer vision tasks. *Advances in Neural Information Processing Systems*, 36, 2024.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial networks. *CoRR*, abs/1406.2661, 2014. URL http://arxiv.org/abs/1406.2661.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6572.

Priya Goyal, Mathilde Caron, Benjamin Lefaudeux, Min Xu, Pengchao Wang, Vivek Pai, Mannat Singh, Vitaliy Liptchinsky, Ishan Misra, Armand Joulin, et al. Self-supervised pretraining of visual features in the wild. *arXiv preprint arXiv:2103.01988*, 2021a.

Priya Goyal, Quentin Duval, Jeremy Reizenstein, Matthew Leavitt, Min Xu, Benjamin Lefaudeux, Mannat Singh, Vinicius Reis, Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Ishan Misra. Vissl. https://github.com/facebookresearch/vissl, 2021b.

Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):87–110, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 770–778. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.90. URL https://doi.org/10.1109/CVPR.2016.90.

Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. In Jennifer G. Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2142–2151. PMLR, 2018. URL http://proceedings.mlr.press/v80/ilyas18a.html.

Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):4037–4058, 2020.

Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(11):4037–4058, 2021. doi: 10.1109/TPAMI.2020.2992393. URL https://doi.org/10.1109/TPAMI.2020.2992393.

Ziv Katzir and Yuval Elovici. Who's afraid of adversarial transferability? *CoRR*, abs/2105.00433, 2021. URL https://arxiv.org/abs/2105.00433.

Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In *European conference on computer vision*, pp. 491–507. Springer, 2020.

Alex Krizhevsky. Learning multiple layers of features from tiny images. *preprint xxxx*, pp. 32–33, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In Peter L. Bartlett, Fernando C. N. Pereira, Christopher J. C. Burges, Léon Bottou, and Kilian Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*, pp. 1106–1114, 2012. URL https://proceedings.neurips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html.

Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=HJGU3Rodl.

Qizhang Li, Yiwen Guo, and Hao Chen. Yet another intermediate-level attack. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (eds.), *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XVI*, volume 12361 of *Lecture Notes in Computer Science*, pp. 241–257. Springer, 2020. doi: 10.1007/978-3-030-58517-4_15. URL https://doi.org/10.1007/978-3-030-58517-4_15.

Qizhang Li, Yiwen Guo, Wangmeng Zuo, and Hao Chen. Improving adversarial transferability via intermediate-level perturbation decay. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/67b2e2e895380fa6acd537c2894e490e-Abstract-Conference.html.

Yandong Li, Lijun Li, Liqiang Wang, Tong Zhang, and Boqing Gong. NATTACK: learning the distributions of adversarial examples for an improved black-box attack on deep neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3866–3876. PMLR, 2019. URL http://proceedings.mlr.press/v97/li19g.html.

Zhiwei Li, Min Ren, Qi Li, Fangling Jiang, and Zhenan Sun. Improving transferability of adversarial samples via critical region-oriented feature-level attack. *IEEE Trans. Inf. Forensics Secur.*, 19:6650–6664, 2024. doi: 10.1109/TIFS.2024.3404857. URL https://doi.org/10.1109/TIFS.2024.3404857.

Nicholas A. Lord, Romain Müller, and Luca Bertinetto. Attacking deep networks with surrogate-based adversarial black-box methods is easy. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=Zf4ZdI4OQPV.

Wenshuo Ma, Yidong Li, Xiaofeng Jia, and Wei Xu. Transferable adversarial attack for both vision transformers and convolutional networks via momentum integrated gradients. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 4607–4616. IEEE, 2023. doi: 10.1109/ICCV51070.2023.00427. URL https://doi.org/10.1109/ICCV51070.2023.00427.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

María Luisa Menéndez, JA Pardo, L Pardo, and MC Pardo. The jensen-shannon divergence. *Journal of the Franklin Institute*, 334(2):307–318, 1997.

Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 6706–6716. Computer Vision Foundation / IEEE, 2020. doi: 10.1109/CVPR42600.2020.00674. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Misra_Self-Supervised_Learning_of_Pretext-Invariant_Representations_CVPR_2020_paper.html.

Seungyong Moon, Gaon An, and Hyun Oh Song. Parsimonious black-box adversarial attacks via efficient combinatorial optimization. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4636–4645. PMLR, 2019. URL http://proceedings.mlr.press/v97/moon19a.html.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: A simple and accurate method to fool deep neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 2574–2582. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.282. URL https://doi.org/10.1109/CVPR.2016.282.

Nina Narodytska and Shiva Prasad Kasiviswanathan. Simple black-box adversarial perturbations for deep networks. *CoRR*, abs/1612.06299, 2016. URL http://arxiv.org/abs/1612.06299.

Muzammal Naseer, Salman H. Khan, Munawar Hayat, Fahad Shahbaz Khan, and Fatih Porikli. A self-supervised approach for adversarial robustness. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 259–268. Computer Vision Foundation / IEEE, 2020a. doi: 10.1109/CVPR42600.2020.00034. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Naseer_A_Self-supervised_Approach_for_Adversarial_Robustness_CVPR_2020_paper.html.

Muzammal Naseer, Salman Hameed Khan, Munawar Hayat, Fahad Shahbaz Khan, and Fatih Murat Porikli. A self-supervised approach for adversarial robustness. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 259–268, 2020b. URL https://api.semanticscholar.org/CorpusID:219559020.

Muzammal Naseer, Kanchana Ranasinghe, Salman Khan, Fahad Shahbaz Khan, and Fatih Porikli. On improving adversarial transferability of vision transformers. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=D6nH3719vZy.

Alejandro Newell and Jia Deng. How useful is self-supervised pretraining for visual tasks? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7345–7354, 2020.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pp. 722–729. IEEE, 2008.

Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (eds.), *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI*, volume 9910 of *Lecture Notes in Computer Science*, pp. 69–84. Springer, 2016. doi: 10.1007/978-3-319-46466-4_5. URL https://doi.org/10.1007/978-3-319-46466-4_5.

Nicolas Papernot, Patrick D. McDaniel, Ian J. Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In Ramesh Karri, Ozgur Sinanoglu, Ahmad-Reza Sadeghi, and Xun Yi (eds.), *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security, AsiaCCS 2017, Abu Dhabi, United Arab Emirates, April 2-6, 2017*, pp. 506–519. ACM, 2017. doi: 10.1145/3052973.3053009. URL https://doi.org/10.1145/3052973.3053009.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 3498–3505. IEEE, 2012.

V Porkodi, M Sivaram, Amin Salih Mohammed, and V Manikandan. Survey on white-box attacks and solutions. *Asian Journal of Computer Science and Technology*, 7(3):28–32, 2018.

Omid Poursaeed, Isay Katsman, Bicheng Gao, and Serge J. Belongie. Generative adversarial perturbations. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 4422–4431. Computer Vision Foundation / IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00465. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Poursaeed_Generative_Adversarial_Perturbations_CVPR_2018_paper.html.

Yunxiao Qin, Yuanhao Xiong, Jinfeng Yi, and Cho-Jui Hsieh. Training meta-surrogate model for transferable adversarial attack. In Brian Williams, Yiling Chen, and Jennifer Neville (eds.), *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pp. 9516–9524. AAAI Press, 2023. doi: 10.1609/AAAI.V37I8.26139. URL https://doi.org/10.1609/aaai.v37i8.26139.

Vivek B. S., Konda Reddy Mopuri, and R. Venkatesh Babu. Gray-box adversarial training. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (eds.), *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XV*, volume 11219 of *Lecture Notes in Computer Science*, pp. 213–228. Springer, 2018. doi: 10.1007/978-3-030-01267-0_13. URL https://doi.org/10.1007/978-3-030-01267-0_13.

Lars St, Svante Wold, et al. Analysis of variance (anova). *Chemometrics and intelligent laboratory systems*, 6(4):259–272, 1989.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In Yoshua Bengio and Yann LeCun (eds.), *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. URL http://arxiv.org/abs/1312.6199.

Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. YFCC100M: the new data in multimedia research. *Commun. ACM*, 59(2):64–73, 2016. doi: 10.1145/2812802. URL https://doi.org/10.1145/2812802.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10347–10357. PMLR, 2021. URL http://proceedings.mlr.press/v139/touvron21a.html.

Futa Waseda, Sosuke Nishikawa, Trung-Nghia Le, Huy H. Nguyen, and Isao Echizen. Closer look at the transferability of adversarial examples: How they fool different models differently. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2023, Waikoloa, HI, USA, January 2-7, 2023*, pp. 1360–1368. IEEE, 2023. doi: 10.1109/WACV56688.2023.00141. URL https://doi.org/10.1109/WACV56688.2023.00141.

Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 3733–3742. Computer Vision Foundation / IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00393. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Wu_Unsupervised_Feature_Learning_CVPR_2018_paper.html.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.

Jianping Zhang, Weibin Wu, Jen-tse Huang, Yizhan Huang, Wenxuan Wang, Yuxin Su, and Michael R. Lyu. Improving adversarial transferability via neuron attribution-based attacks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 14973–14982. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01457. URL https://doi.org/10.1109/CVPR52688.2022.01457.

Richard Zhang, Phillip Isola, and Alexei A. Efros. Colorful image colorization. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (eds.), *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III*, volume 9907 of *Lecture Notes in Computer Science*, pp. 649–666. Springer, 2016. doi: 10.1007/978-3-319-46487-9_40. URL https://doi.org/10.1007/978-3-319-46487-9_40.

Yao Zhu, Yuefeng Chen, Xiaodan Li, Kejiang Chen, Yuan He, Xiang Tian, Bolun Zheng, Yaowu Chen, and Qingming Huang. Toward understanding and boosting adversarial transferability from a distribution perspective. *IEEE Trans. Image Process.*, 31:6487–6501, 2022. doi: 10.1109/TIP.2022.3211736. URL https://doi.org/10.1109/TIP.2022.3211736.

Adrian Ziegler and Yuki M Asano. Self-supervised learning of object parts for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14502–14511, 2022.

Model Families	CIFAR10	CIFAR100	Oxford Flowers	Oxford Pets
AlexNet (Colorization, IN1K)	88.97	98.96	24.91	49.94
AlexNet (Colorization, IN22K)	89.56	98.92	25.19	50.06
AlexNet (Colorization, YFCC100M)	87.84	98.55	24.91	49.96
AlexNet (Jigsaw, IN1K)	53.25	74.03	26.96	45.38
AlexNet (Jigsaw, IN22K)	53.06	73.76	30.61	49.86
AlexNet (DeepCluster V2)	49.59	64.38	27.15	44.52
ResNet-50 (Jigsaw, IN22K)	61.03	81.81	26.37	47.28
ResNet-50 (Colorization, IN1K)	89.86	98.07	24.91	50.12
ResNet-50 (Colorization, IN22K)	88.99	97.89	27.01	50.00
ResNet-50 (Jigsaw, IN1K)	56.34	80.01	25.46	48.12
ResNet-50 (Jigsaw, IN22K)	54.48	75.08	26.79	47.75
ResNet-50 (RotNet, IN1K)	47.71	72.61	37.86	45.69
ResNet-50 (Jigsaw, IN1K)	58.02	78.32	26.17	48.06
ResNet-50 (NPID)	58.37	80.39	49.77	48.42
ResNet-50 (PIRL)	58.80	84.12	34.03	44.10
ResNet-101 (SimCLR)	55.09	70.34	28.54	47.12
ResNet-50 (SimCLR)	51.57	65.91	30.26	44.12
ResNet-50 (SwAV, 400ep)	48.63	68.46	28.79	44.33
ResNet-50 (SwAV, 800ep)	50.23	67.89	27.73	45.33
DeiT-Small (DINO)	63.37	85.08	26.56	47.26
XCiT-Small (DINO)	49.46	64.84	27.19	46.76

Table 2: Adversarial Transferability Averaged for each dataset per model architecture type

A Experimental Detials

A.1 Model variations and Adversarial Transferability

The adversarial transferability for each type of model can be seen in Table 2. The complete set of model variations used throughout the experimentations can be observed in Table 3.

A.2 Computational Resources

All experiments were conducted using two compute nodes, each equipped with 8 NVIDIA A100 GPUs (80 GB memory per GPU), resulting in a total of 16 GPUs. Each node was powered 96 vCPUs (Intel Xeon Platinum) and 400 GB of RAM. Training all 352 model variations required approximately 3200 GPU-hours. The adversarial evaluation phase—including proxy attack generation, backbone attacks, and high-budget PGD experiments—required an additional 1800 GPU-hours. To ensure consistency, we fixed all random seeds to 42 across all runs, including for NumPy, PyTorch, and Python's built-in random module. Model tuning configurations, checkpoints, logs, and attack results were stored for full reproducibility.

SSL Method	Pretraining Dataset	Architecture
Colorization (Zhang et al., 2016)		
Colorization	YFCC100M	AlexNet
Colorization	ImageNet-1K	AlexNet
Colorization	ImageNet-1K	ResNet-50
Colorization	ImageNet-21K	AlexNet
Colorization	ImageNet-21K	ResNet-50
Jigsaw Puzzle (Noroozi & Favaro, 2016)		
Jigsaw Puzzle	ImageNet-21K	ResNet-50
Jigsaw Puzzle	ImageNet-1K	ResNet-50
Jigsaw Puzzle	ImageNet-21K	ResNet-50
Jigsaw Puzzle	ImageNet-21K	AlexNet
Jigsaw Puzzle	ImageNet-1K	AlexNet
Jigsaw Puzzle	ImageNet-1K	ResNet-50
PIRL (Jigsaw-based) (Misra & van der Maaten, 2020)		
PIRL	ImageNet-1K	ResNet-50
Rotation Prediction (Gidaris et al., 2018)		
RotNet	ImageNet-1K	ResNet-50
DINO (Caron et al., 2021)		
DINO	ImageNet-1K	DeiT-Small
DINO	ImageNet-1K	XCiT-Small
SimCLR (Chen et al., 2020a)		
SimCLR	ImageNet-1K	ResNet-50
SimCLR	ImageNet-1K	ResNet-101
SwAV (Caron et al., 2020)		
SwAV	ImageNet-1K	ResNet-50
SwAV	ImageNet-1K	ResNet-50
DeepCluster V2 (Caron et al., 2018)		
DeepCluster V2	ImageNet-1K	AlexNet
Instance Discrimination (NPID) (Wu et al., 2018)		
NPID	ImageNet-1K	ResNet-50

Table 3: Summary of Self-Supervised Learning Methods, Pretraining Datasets, and Architectures used in our study.