

Adversarial Robustness of Discriminative Self-Supervised Learning in Vision

Ömer Veysel Çağatan^{1,2} ¹Ömer Faruk Tal ¹M. Emre Gürsoy

¹Department of Computer Engineering, Koç University

²KUIS AI Center, Koç University

ocagatan19@ku.edu.tr, otal19@ku.edu.tr, emregursoy@ku.edu.tr

Abstract

Self-supervised learning (SSL) has advanced significantly in visual representation learning, yet comprehensive evaluations of its adversarial robustness remain limited. In this study, we evaluate the adversarial robustness of seven discriminative self-supervised models and one supervised model across diverse tasks, including ImageNet classification, transfer learning, segmentation, and detection. Our findings suggest that discriminative SSL models generally exhibit better robustness to adversarial attacks compared to their supervised counterpart on ImageNet, with this advantage extending to transfer learning when using linear evaluation. However, when fine-tuning is applied, the robustness gap between SSL and supervised models narrows considerably. Similarly, this robustness advantage diminishes in segmentation and detection tasks. We also investigate how various factors might influence adversarial robustness, including architectural choices, training duration, data augmentations, and batch sizes. Our analysis contributes to the ongoing exploration of adversarial robustness in visual self-supervised representation systems.

1. Introduction

Self-supervised learning (SSL) [4], particularly discriminative approaches, has emerged as a foundational method for training models with remarkable capabilities in areas such as language [80], vision [65], and decision-making [50]. As these models become increasingly widespread and integrated into various applications, ensuring their reliability and safety has become a critical concern [6, 7].

One particular challenge is the surprising vulnerability of deep learning models to adversarial examples, where slight input alterations can significantly impact model performance [35, 78]. This phenomenon has sparked significant debate, seeking to understand and mitigate these vulnerabilities [3, 32, 75, 76, 79, 83–85]. One prominent theory [47] suggests that adversarial examples arise from the model’s sensitivity to non-robust features in the input data.

According to this view, both robust (stable) and non-robust (vulnerable) features contribute to classification, with adversarial attacks manipulating the latter to cause misclassification. However, this theory, developed primarily in the context of supervised learning, faces challenges when extended to other self-supervised paradigms. [56] indicates that non-robust features are less effective in SSL methods such as contrastive learning [16], masked image modeling [40], or diffusion models [43]. This discrepancy suggests that non-robust features may lack the transferability across learning paradigms that robust or natural features possess. Thus, it becomes essential to systematically evaluate and compare how different SSL approaches respond to adversarial attacks, particularly given the theoretical evidence suggesting their feature representations may differ fundamentally from supervised models.

These theoretical insights into how adversarial examples affect different learning paradigms highlight several critical gaps in our understanding of SSL’s adversarial robustness. Notwithstanding the progress made in understanding the adversarial robustness of SSL, particularly contrastive learning, which we extensively discuss in Section 2, several key questions remain unresolved. First, with the wide variety of self-supervised representations available, employing different pretext tasks and data augmentations, which approaches demonstrate the greatest adversarial robustness? This remains unclear since most methods don’t provide any results on adversarial robustness unless it is a specific focus of the proposed approach. Secondly, robustness is typically assessed by the model’s accuracy on the pretraining dataset. Still, its adversarial impact on transfer learning or downstream tasks like detection and segmentation has not been thoroughly investigated [52].

The choice of model architecture also raises questions about robustness. Standard vision SSL pretraining typically utilizes a ResNet [38] as the backbone, but more recently, larger and more powerful models [12, 20, 65] have been developed using vision transformers [28]. This leads to the question: Which architecture demonstrates greater robustness under the same SSL objective and with comparable

parameter sizes?

Another factor to consider is the training duration. State-of-the-art SSL models are trained for longer durations compared to their supervised counterparts. Several studies indicate that this extended training consistently enhances performance, raising the question of whether this might compromise the models' adversarial robustness.

While previous work has examined aspects of adversarial robustness in SSL, our study provides the first comprehensive cross-model comparison across multiple tasks, architectures, and training regimes. We assess seven different SSL models (Barlow Twins [89], BYOL [36], DINO [12], MoCoV3 [20], SimCLR [16], SwAV [11], and VICReg [5]) alongside a supervised model against various adversarial attacks on ImageNet [74] and nine other image-recognition datasets. We also evaluate their robustness in segmentation and detection tasks. Our investigation addresses the following key questions:

1. How does the adversarial robustness of various SSL models compare to that of supervised models on ImageNet?

SSL models consistently demonstrate greater adversarial robustness than supervised models on ImageNet. Non-contrastive methods show particular resilience against IAA [13] attacks, while all SSL approaches exhibit strong resistance to UAP [14], with MoCoV3 demonstrating the strongest overall performance.

2. Does SSL robustness transfer to downstream tasks like transfer learning, segmentation, and detection?

The robustness advantages transfer effectively to classification tasks via linear probing and fine-tuning. However, in segmentation and detection, all models exhibit similar vulnerability regardless of pretraining methodology, suggesting task-specific architectural components may override backbone robustness properties.

3. How does model architecture influence adversarial robustness under the same SSL objective?

Architecture impact is highly objective dependent. MoCoV3 shows reduced robustness with Vision Transformers, whereas DINO demonstrates improved performance with ViT compared to ResNet, challenging the notion that architectural effects are uniform across SSL paradigms. Additionally, we have evaluated DINOv2 [65] and MAE [40] that do not demonstrate a consistently high or low level of robustness.

4. Does extending training duration compromise adversarial robustness in SSL models?

Extended training either maintains or slightly enhances adversarial performance. For UAP attacks, performance improves meaningfully after more epochs in both SwAV and MoCoV3, indicating that longer training periods do not create a performance-robustness tradeoff.

2. Related Work

Self Supervised Learning Self-supervised learning (SSL) seeks to extract meaningful and general representations from unlabeled data by leveraging pretext tasks. These tasks can vary, such as predicting the next word [71] or neighboring words [26] in a text, reconstructing masked sections of an image [40], or ensuring that two different perspectives of the same image result in similar visual representations [16].

Avoiding collapse is a key challenge in SSL for computer vision, and various methods can be classified based on how they address this issue. Contrastive approaches like SimCLR [16] and MoCo [19, 20, 39] use an objective that pushes apart representations of different inputs (negative samples) while bringing together those of the same input (positive samples). The performance and scalability of these methods heavily depend on the number and selection of negative samples. In another category, distillation methods such as BYOL [36], SimSiam [18], and DINO [12], prevent collapse by introducing asymmetry between different encoder branches and employing algorithmic adjustments. Additional SSL techniques, including DeepCluster [10], SeLa [1], and SwAV [11], enforce a clustering structure in the feature space to avoid constant representations. Meanwhile, methods like Barlow Twins [89], Whitening MSE (W-MSE) [30], VICReg [5], CorInfoMax [67] prevent collapse by using feature decorrelation.

Adversarial Self-Supervised Learning While self-supervised learning (SSL) has outperformed supervised training [16], numerous studies highlight that contrastive learning remains susceptible to adversarial attacks when transferring the learned features to downstream classification tasks [42, 49]. To improve the robustness of contrastive learning, adversarial training has been adapted to self-supervised settings. In the absence of labels, adversarial examples are generated by maximizing the contrastive loss with respect to all input samples. Several prior works, such as ACL [48], RoCL [49], and CLAE [42], adopt this approach. Additionally, ACL incorporates the dual-BN technique [87] to further enhance performance. DeACL [90] introduces a two-stage approach, distilling a standard pre-trained encoder through adversarial training. Nguyen et al. [64] establishes an upper bound on the adversarial loss of a prediction model, which is based on the learned representations, for any downstream task. This upper bound is determined using the model's loss on clean data and a robustness regularization term, which helps make the prediction model more resistant to adversarial attacks. [37] demonstrates that adversarial sensitivity stems from the uniform distribution of data representations on a unit hypersphere in the representation space. The presence of false negative pairs during training contributes to this effect, increasing the model's vulnerability to input perturbations.

Although self-supervised adversarial training has made

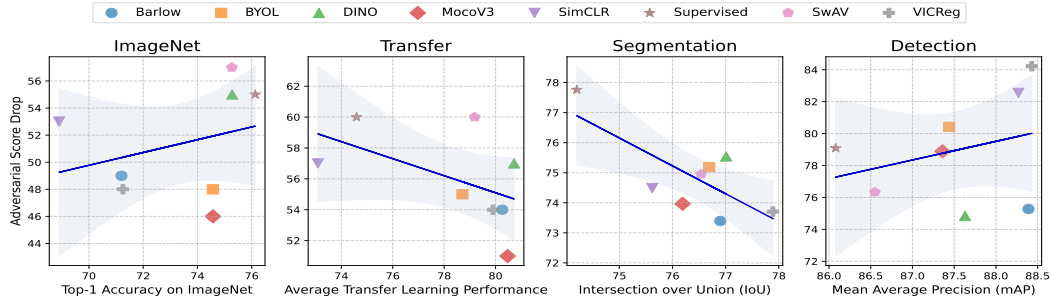


Figure 1. Performance scores for tasks such as ImageNet classification, transfer learning (with linear probing), segmentation, and detection (both with frozen backbones) are shown with the percentage drop in adversarial robustness. The shaded regions indicate the 95% confidence interval around the regression line. Note the consistent pattern of higher robustness (lower percentage drop) among SSL models compared to supervised approaches in classification tasks.

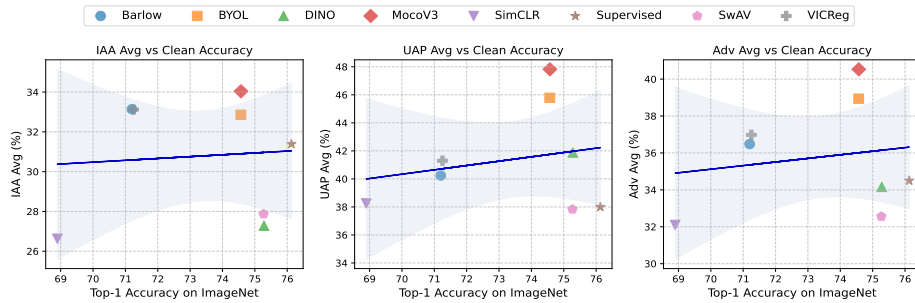


Figure 2. Averaged scores of SSL models on ImageNet across various attack types, including Instance Adversarial Attacks (IAA) and Universal Adversarial Perturbations (UAP). *Adv Avg* refers to the average score across all attacks combined. The shaded regions indicate the 95% confidence interval around the regression line.

progress, it still does not match the performance of supervised methods. Luo et al. [58] suggests that this shortfall is due to data augmentation and proposes a dynamic data augmentation scheduler to achieve comparable results to supervised training. Xu et al. [88] efficiently applies ACL on the ImageNet [74] to obtain a robust representation using robustness-aware core set selection.

Robustness of Self-Supervised Learning

[41] found that incorporating an extra self-supervised task in a multi-task framework can enhance the adversarial robustness of supervised models. In a similar vein, Carmon et al. [9] discovered that using additional unlabeled data also strengthens the model’s adversarial resilience. Furthermore, Chen et al. [17] created robust variants of pretext-based SSL tasks, showing that their integration with robust fine-tuning leads to a notable increase in robustness compared to standard adversarial training. Chhipa et al. [21] demonstrates a clear relationship between the performance of learned representations within SSL paradigms and the severity of distribution shifts and corruptions, and highlights the critical impact of distribution shifts and image corruptions on the performance and resilience of SSL methods. Similarly, Zhong et al. [91] conduct robustness tests to assess

the behavioral differences between contrastive and supervised learning under changes in downstream or pre-training data distributions, while also exploring the effects of data augmentation and feature space characteristics. Kowalczyk et al. [52] conducts a comprehensive empirical evaluation of the adversarial robustness of self-supervised vision encoders across multiple downstream tasks, revealing the need for broader enhancements in encoder robustness. Goldblum et al. [34] benchmarks diverse pretrained models across multiple computer vision tasks, finding that supervised convolutional neural networks still outperform newer architectures on most metrics, while revealing self-supervised learning backbones show competitive potential when compared under equivalent conditions.

Unlike prior studies that primarily focus on individual SSL methods, specific tasks, or limited adversarial scenarios, our work provides a comprehensive, unified benchmark across multiple SSL paradigms, architectures, and tasks—including classification, transfer learning, segmentation, and detection—under a diverse set of adversarial attacks, offering a broader and deeper understanding of adversarial robustness in SSL.

3. Experimental Setup

3.1. SSL Models

While numerous SSL approaches have been proposed [66], we focus on a subset of widely used models due to computational constraints: Barlow Twins [89], BYOL [36], DINO [12], MoCoV3 [20], SimCLR [16], SwAV [11], and VICReg [5]. We primarily use ResNet50 [38] backbones, as most SSL checkpoints are released in this format, with DINO and MoCoV3 also offering ViT [28] variants. For BYOL, DINO, MoCoV3, SimCLR, and SwAV, complete model checkpoints were available. In contrast, Barlow Twins and VICReg only provided backbone weights, requiring linear evaluation via official code, which led to a slight drop in performance. For comparison, we also include a supervised ResNet50 baseline from PyTorch [68].

3.2. ImageNet and Transfer Learning

We use the benchmark suite introduced in the transfer learning study [45], which encompasses the target datasets like FGVC Aircraft [60], Caltech-101 [33], Stanford Cars [53], CIFAR 10 [54], CIFAR 100 [54], DTD [22], Oxford 102 Flowers [22], and Food-101 [8]. We follow Ericsson et al. [29] for both linear evaluation and fine-tuning of these datasets. We prioritized linear evaluation in our analysis as the backbone remains frozen during this process, allowing for a more equitable comparison of objectives within this setup. We apply the same adversarial techniques to ImageNet and transfer learning: Instance Adversarial Attacks (IAA) and Universal Adversarial Perturbations (UAP). In brief, instance-based methods generate unique perturbations for each image, while UAP creates a single perturbation that applies across the entire dataset. Comprehensive details and categorizations of all attack methods—such as white-box, black-box, gradient-free, among others—are provided in Section 1 of the Supp Mat..

3.3. Segmentation

For segmentation, we use the Pascal VOC 2012 dataset [31] and CityScapes [23] dataset, training a DeepLabV3+ model [15]. To conduct the attacks, we follow the setup from Rony et al. [73], utilizing Alma [73], Asma [73], DAG [86], DDN [73], FGSM [35], FMN [70], and PGD [59]. While our primary metric is the mean Intersection Over Union (mIOU), we also report the Attack Pixel Success Rate (APSR) introduced by [73]. Although our main focus is on using a frozen backbone, we also perform training following the standard procedure. Our CityScapes results include only the frozen backbone approach, while our Pascal VOC results include both frozen and unfrozen backbone configurations.

Table 1. Performance of various models on ImageNet, Transfer Learning, Segmentation (Pascal VOC), and Detection (INRIA Person) tasks with frozen backbones, showing original (Orig.) and adversarial (Adv.) scores with performance drops in red.

Model	ImageNet		Transfer Learning		Segmentation		Detection	
	Orig.	Adv.	Orig.	Adv.	Orig.	Adv.	Orig.	Adv.
Barlow Twins	71.2	36.5 ↓49%	80.3	37.8 ↓54%	76.9	20.5 ↓73%	88.4	21.9 ↓75%
BYOL	74.6	38.9 ↓48%	78.7	36.6 ↓55%	76.7	19.0 ↓75%	87.4	17.3 ↓80%
DINO	75.3	34.2 ↓55%	80.7	35.6 ↓57%	77.0	18.9 ↓76%	87.6	22.0 ↓75%
MoCoV3	74.6	40.5 ↓46%	80.5	40.3 ↓51%	76.2	19.9 ↓74%	87.3	18.5 ↓79%
SimCLR	68.9	32.1 ↓53%	73.1	33.0 ↓57%	75.6	19.3 ↓74%	87.7	13.1 ↓85%
Supervised	76.1	34.5 ↓55%	74.6	31.2 ↓60%	74.2	16.5 ↓78%	86.1	18.0 ↓79%
SwAV	75.3	32.6 ↓57%	79.2	32.4 ↓60%	76.5	19.2 ↓75%	86.6	20.5 ↓76%
VICReg	71.3	37.0 ↓48%	79.9	37.7 ↓54%	77.9	20.5 ↓74%	88.4	14.0 ↓84%

3.4. Detection

For object detection, we utilized the INRIA Person [24] and CoCo [57] datasets, and trained a Faster R-CNN [72]. To perform adversarial attacks, we followed the setup described by [44], employing the Transfer-based Self-Ensemble Attack (T-SEA). The T-SEA attack can be deployed using various methods and optimizers. In our experiments, we employed BIM [44], MIM [27], PGD [59], and Optim [44] methods. Additionally, we explored simpler methods that rely on common optimizers, such as Adam [51], SGD, and Nesterov [63]. Throughout our evaluation, we report the mean average precision (mAP) scores as the primary performance metric. While our primary focus was on employing a frozen backbone, we also conducted training experiments following the standard training procedures for comparative analysis. Our COCO results include only the frozen backbone approach.

4. Results and Discussion

In this section, we present our experimental findings on ImageNet, transfer learning, and detection, and discuss each in turn. Figure 1 and table 1 summarize the performance of various SSL models compared to supervised learning across our main evaluation tasks in the frozen backbone setup. While we address the results individually, the full detailed results are provided in Section 4 of the Supp. Mat..

4.1. ImageNet

SSL vs Supervised. Most robustness studies on contrastive learning [42, 48, 49, 64, 87, 90] focus on small datasets like CIFAR10 [54] and primarily evaluate robustness using adversarial attacks such as FGSM [35] and PGD [59]. While computational constraints explain the reluctance to scale to larger datasets like ImageNet [74], many evaluations inadequately incorporate Universal Adversarial Perturbations (UAP). Our findings, as shown in Figure 2 contradict previous research by Gupta et al. [37]. Under IAA, MoCoV3 demonstrates the strongest robustness (54% drop), while SimCLR shows the weakest performance (61% drop). For

UAP attacks, MoCoV3 again leads (38% drop), while the supervised model and SwAV both show 50% drops. Notably, our results challenge the conclusion about contrastive vs. non-contrastive methods. MoCoV3, a contrastive model, consistently demonstrates the highest adversarial robustness with ResNet architecture, while DINO, which was claimed to perform better due to its non-contrastive nature, shows the weakest IAA robustness in our ResNet evaluation (64% drop). Our data reveals that Barlow, BYOL, MoCoV3, and VICReg all demonstrate comparable resilience against IAA (around 54% drop), contradicting the simple categorization of robustness based on contrastive versus non-contrastive approaches. Our comprehensive evaluation using diverse attack methods demonstrates that the relationship between self-supervised learning approaches and adversarial robustness is more complex than previously suggested. The full results are in Section 4.1 of the Supp. Mat..

What makes MoCoV3 robust? Although MoCoV3 and SimCLR both utilize the InfoNCE [77, 81] objective, there is a stark contrast in their adversarial robustness. To understand this disparity, we evaluate MoCoV1 [39], MoCoV2 [19], and MoCoV3.

A brief MoCo History. *MoCoV1 introduced a dynamic dictionary with a queue and momentum-updated encoder to maintain consistent negative samples. MoCoV2 enhanced this with a multi-layer projection head and stronger data augmentation. MoCoV3 further evolved by eliminating the memory bank and incorporating a prediction head similar to BYOL and SimSiam [18].*

Our analysis reveals a clear progression in robustness across MoCo versions. While MoCoV1 demonstrates limited resilience with a 71% drop in overall adversarial accuracy, MoCoV2 shows significant improvement (52% drop), and MoCoV3 achieves the strongest performance (46% drop). The most dramatic enhancement occurs in UAP resistance, where MoCoV1’s performance drops by 77%, compared to MoCoV2’s 39% and MoCoV3’s 36%. The substantial improvement from MoCoV1 to MoCoV2 (7 % in clean accuracy, 27 percentage points in UAP robustness) primarily stems from the non-linear projector, with data augmentation providing marginal benefits. While [46] suggests non-linear projectors aren’t always essential, our results indicate they significantly enhance both performance and adversarial robustness. MoCoV3’s superior performance over MoCoV2 (an additional 7% in clean accuracy and 7 percentage points in UAP robustness) can be attributed to its prediction head and larger batch size. Unlike the transition from V1 to V2, MoCoV3 shows substantial improvements in both IAA (34% vs 25%) and UAP (48% vs 41%), highlighting the prediction head’s critical role in enhancing overall robustness. Our findings suggest that momentum, a common feature in robust models like MoCoV3 and BYOL, significantly contributes to adversar-

ial resilience, while MoCoV2’s performance more closely resembles that of SimCLR, which lacks this feature. The full results are in Section 4.6 of the supplementary material.

Augmentations vs Algorithms.

Morningstar et al. [61] challenges the notion that SSL progress is primarily driven by algorithmic advancements, suggesting instead that augmentation diversity, along with data and model scale, play more critical roles. Their analysis argues that many algorithmic improvements, such as prediction networks or new loss functions, had minimal impact on downstream task performance compared to stronger augmentation techniques.

Our comprehensive analysis of the MoCo family evolution provides a more nuanced perspective on this debate. The substantial progression in adversarial robustness from MoCoV1 to MoCoV2 and further to MoCoV3 suggests that architectural innovations like non-linear projectors and prediction heads significantly impact robustness. MoCoV2’s dramatic improvement over MoCoV1, particularly in UAP performance, indicates that the multi-layer projection head provides substantial benefits beyond mere augmentation changes. Similarly, MoCoV3’s further enhancements in both IAA and UAP performance relative to MoCoV2 highlight the crucial role of the prediction head in overall robustness.

While our comparison lacks perfectly controlled baselines for augmentations across different objectives (with slight variations in augmentation between MoCoV2 and V3), this limitation stems from the unavailability of public checkpoints rather than our experimental design. Despite this constraint, the consistent improvements in adversarial performance strongly suggest that algorithmic innovations significantly contribute to adversarial robustness. Importantly, our findings demonstrate that higher clean accuracy doesn’t automatically translate to improved robustness on ImageNet.

Although the checkpoints from [61] are not publicly available, we conducted an ablation study on augmentation types and batch sizes using BYOL which is the only model in our SSL pool that includes configurations with varied augmentations and batch sizes. For this analysis, we consider four distinct models: BYOL-NC (without color distortions), BYOL-CC (with only color and crop augmentations), and the standard BYOL models with batch sizes of 128 (BYOL-128) and 512 (BYOL-512).

We observe that batch size and augmentation choices have varying effects on the adversarial robustness of BYOL variants. For IAA robustness, both BYOL-NC and BYOL-CC show similar performance (60% drop), while BYOL-128, BYOL-512, and standard BYOL demonstrate slightly better resilience (57%, 56%, and 56% drops, respectively). These modest differences suggest that batch size and augmentation have a limited impact on instance-level attack ro-

bustness.

For UAP attacks, we see more substantial variations: BYOL-128 (41% drop) and BYOL-512 (43% drop) perform comparably, while BYOL-NC and BYOL-CC show notably weaker performance (46% and 48% drops). Standard BYOL demonstrates a 39% drop, which is marginally better than the batch size variants but substantially better than the limited-augmentation models. The 7-9 percentage point difference between standard BYOL and the limited-augmentation variants suggests that comprehensive augmentation strategies may contribute to improved UAP robustness.

Interestingly, we note that BYOL-CC performs slightly worse than BYOL-NC despite having more augmentations, though this difference is too small to draw meaningful conclusions. Similarly, the differences between BYOL-128, BYOL-512, and standard BYOL in UAP robustness are relatively minor and should be interpreted cautiously. Our results indicate that the relationship between augmentation and adversarial robustness is complex and that meaningful improvements likely require more than just incremental changes to batch size or augmentation strategies. While our data hints at potential benefits from comprehensive augmentation for UAP robustness, the effects are modest and require further investigation with more controlled experiments. The full results are in Section 4.7 of the supplementary material.

ResNet vs ViT in Adversarial Robustness. While ViTs are generally seen as more robust than CNNs [62], Bai et al. [2], Pinto et al. [69] demonstrate that with the right training methods, CNNs [55] can achieve comparable robustness. Despite ViT’s success [12, 20, 25, 28, 65], most SSL methods still use ResNet for validation. We examine MoCoV3 and DINO, as they are the only models that include ViT training, with our analysis covering both standard ViTs comparable to ResNet50 and larger ViT-B variants with approximately 4x more parameters.

Our results reveal notable architectural differences. DINO performs significantly better with ViT architectures than ResNet, with DINO-ViT-B achieving 45.19% adversarial average (42% drop from clean accuracy) compared to DINO-ResNet’s 55% drop. In contrast, MoCo struggles with transformer architectures, with MoCo-ViT showing poor performance at a 61% drop, significantly worse than MoCoV3-ResNet at a 46% drop. Even MoCoV3-ViT-B 51% drop underperforms its ResNet counterpart despite having 4x more parameters.

These differences are especially evident in UAP results, where DINO-ViT-B shows greater resilience (38% drop) than DINO-ResNet (44% drop), while MoCo-ViT performs poorly (26.31%, 64% drop) compared to MoCoV3-ResNet (36% drop). These findings suggest that the interaction between self-supervised approaches and model architectures

significantly impacts adversarial robustness, with DINO benefiting from the ViT architecture while MoCo struggles with transformer models. The full results are in Section 4.6 of the Supp. Mat..

Lastly, we have evaluated DINOv2’s [65] small and base and MAE’s [40] base versions. We discuss them in the section 2 in Supp. Mat..

Impact of Training Duration. SSL models tend to demonstrate better performance as training epochs increase [11, 16, 20]. However, due to computational constraints, many models are reported with different numbers of epochs. This prompts the question of whether longer training durations enhance or reduce adversarial robustness. As ViT models do not have checkpoints at various epochs, we focus on ResNet-based SSL models, specifically SwAV and MoCoV3, which offer multiple checkpoints throughout the training process. We find that both SwAV and MoCo show very marginal improvement of about 1% on IAA across various epochs, which is minimal compared to the rise in original accuracy. In contrast, both methods exhibit a modest increase in UAP performance after surpassing 100 epochs, with SwAV improving from 33% to 40% and MoCoV3 from 41% to 48%. Overall, our results suggest that despite differences in reported checkpoints, robustness generally remains largely stable during training for IAA, with slightly more noticeable but still modest gains for UAP. This indicates that training duration has a limited impact on adversarial robustness, even as clean accuracy continues to improve with extended training. The full results are in Section 4.5 of the Supp. Mat.

4.2. Transfer Learning

In this section, we analyze how adversarial robustness transfers from ImageNet pre-training to downstream tasks. We examine whether vulnerability patterns established during pre-training persist when models are evaluated through linear probing or fine-tuning across various ResNet-based self-supervised and supervised learning approaches. This analysis quantifies robustness transfer relationships through both Spearman rank and Pearson correlations.

LINEAR. Our correlation analysis shows strong relationships between ImageNet and linear evaluation vulnerability. For performance drops, both Spearman and Pearson correlations are high across all attack types (Spearman—Adv: 0.93, UAP: 0.83, IAA: 0.79; Pearson—Adv: 0.94, UAP: 0.94, IAA: 0.88). These consistently high correlations suggest that robustness characteristics established during pre-training largely persist through linear probing. The performance drop analysis reveals method-specific robustness characteristics. MoCoV3 consistently exhibits smaller performance drops across both ImageNet (IAA: 54%, UAP: 36%) and linear evaluation (IAA: 56%, UAP: 43%), indicating its contrastive learning approach with mo-

Table 2. Component Contribution Analysis Across Datasets. Metrics include: Original Accuracy (Orig), Adversarial Accuracy (Adv) with relative performance drop percentage, Direction Ratio (D) and Magnitude Ratio (M). Ratio interpretation: Head dominant (<0.67), Balanced (0.67-1.5), Backbone dominant (>1.5).

Model	CIFAR-10				CIFAR-100				ImageNet				CIFAR-10 (FT)				CIFAR-100 (FT)								
	Orig	Adv ↓	D	M	Orig	Adv ↓	D	M	Orig	Adv ↓	D	M	Orig	Adv ↓	D	M	Orig	Adv ↓	D	M					
Barlow	92.3	33.0	↓64%	2.7	3.1	77.9	20.5	↓74%	1.4	5.0	71.2	42.4	↓40%	0.9	9.4	97.1	66.9	↓31%	1.5	2.4	84.6	45.0	↓57%	0.8	2.4
BYOL	93.0	31.0	↓67%	1.6	4.2	78.2	19.01	↓76%	1.07	6.0	74.6	39.4	↓47%	0.9	12.5	96.9	67.0	↓31%	1.2	1.4	83.9	61.2	↓27%	0.7	2.7
DINO	93.9	27.6	↓71%	3.2	3.3	79.7	16.0	↓80%	1.6	6.2	75.3	24.7	↓67%	1.2	14.1	96.9	77.8	↓20%	1.9	2.9	84.7	52.9	↓38%	1.79	4.6
MoCoV3	94.7	33.0	↓65%	2.5	1.7	80.2	19.3	↓76%	1.1	4.5	74.6	42.7	↓43%	0.7	10.9	96.9	72.0	↓26%	1.1	1.3	84.5	62.6	↓26%	0.7	2.6
SimCLR	91.0	37.9	↓58%	1.5	1.9	73.0	19.5	↓73%	1.3	3.5	68.9	24.3	↓65%	1.0	7.7	97.2	67.3	↓31%	1.0	1.9	84.4	43.0	↓50%	0.6	2.0
Supervised	91.4	42.8	↓53%	1.8	0.7	73.9	24.5	↓67%	1.3	1.5	76.1	38.8	↓49%	1.4	3.0	96.2	62.3	↓35%	2.0	0.5	82.6	59.3	↓28%	1.3	1.1
SwAV	93.9	19.4	↓79%	3.1	3.0	79.4	11.1	↓86%	1.8	6.4	75.3	24.7	↓67%	1.2	17.4	96.8	79.9	↓17%	1.7	2.7	84.4	54.6	↓35%	1.0	3.0
VICReg	92.8	33.0	↓64%	2.7	3.1	77.8	22.3	↓71%	1.3	4.3	71.3	42.4	↓40%	0.9	9.3	97.1	68.6	↓29%	1.4	2.4	84.3	43.4	↓48%	0.8	2.4

mentum encoders develops features with inherently better transferable robustness properties. In contrast, supervised pre-training and clustering-based approaches like SwAV experience more severe drops in both settings, suggesting these methods may create more brittle representations.

The high correlation in adversarial vulnerability between ImageNet and linear evaluation indicates that linear probing provides a reliable assessment of a pre-trained model’s downstream robustness characteristics without requiring extensive adaptation. This finding has practical implications for model selection, suggesting that robustness evaluations on ImageNet can effectively predict linear transfer performance. The full results are in Section 4.8 of the Supp. Mat..

FINETUNE. Fine-tuning reveals distinct transfer patterns by attack type. For IAA, the correlation between ImageNet and fine-tuned vulnerability nearly vanishes for performance drops (Spearman: 0.12, Pearson: 0.36), indicating that fine-tuning substantially reshapes defense against instance-specific attacks. Conversely, UAP vulnerability correlation remains high (Spearman: 0.88, Pearson: 0.93), suggesting that susceptibility to universal perturbations is more persistently encoded in the network regardless of parameter adaptation. This attack-dependent correlation disparity reveals that universal perturbation vulnerabilities, which exploit systematic weaknesses across the feature space, are deeply encoded in the network’s architecture and learning approach, while instance-specific vulnerabilities are more malleable through fine-tuning. The absolute magnitude of performance drops diminishes considerably after fine-tuning across all methods (IAA: 49% vs. 60% in linear, UAP: 43% vs. 49% in linear), highlighting fine-tuning’s effectiveness in mitigating vulnerability. MoCoV3 continues to demonstrate superior robustness with the smallest drops (IAA: 45%, UAP: 38%), while BYOL shows dramatic improvement from ImageNet to fine-tuning for IAA attacks (from 56% to 47%). These findings suggest different mechanisms for robustness transfer: UAP vulnerability appears tied to fundamental architectural and al-

gorithmic properties that persist across transfer paradigms, while instance-specific attack vulnerability depends more on the fine-tuning process than initial representation properties. This distinction may be valuable for practical applications, suggesting that pre-training method selection strongly impacts universal attack robustness, while defense against instance-specific attacks can be substantially improved through appropriate fine-tuning strategies. The full results are in Section 4.9 of the Supp. Mat..

Attributing Adversarial Vulnerability. Having established the correlations between ImageNet and transfer learning robustness patterns, we now seek to understand the underlying mechanisms causing these disparities across models and training paradigms. Specifically, we investigate whether adversarial vulnerability stems primarily from the backbone feature extractor or the classification head and how this attribution changes between linear probing and fine-tuning.

Our analysis investigates the relationship between adversarial robustness and component-specific vulnerability under FGSM₁ attack (detail in Supp. Mat.), focusing on two metrics: the Direction Ratio $D = \frac{\text{mean}(1 - \cos(\mathbf{l}, \mathbf{l}_{adv}))}{\text{mean}(1 - \cos(\mathbf{f}, \mathbf{f}_{adv}))}$, which compares directional shifts in logits (\mathbf{l}) versus backbone features (\mathbf{f}), and the Magnitude Ratio $M = \frac{\text{mean}(\|\mathbf{l} - \mathbf{l}_{adv}\|_2)}{\text{mean}(\|\mathbf{f} - \mathbf{f}_{adv}\|_2)}$, quantifying relative sensitivity to perturbation magnitudes. We fit linear regression models using these ratios to predict adversarial accuracy drop across models.

For probed models, Direction Ratio strongly correlates with vulnerability ($R^2 = 0.56-0.85$), suggesting that directional instability in the head dominates FGSM robustness. SwAV, with $D = 3.1$ on CIFAR-10, suffers a 79% accuracy drop, while SimCLR ($D = 1.5$) drops 58%. This may arise from the linear head’s limited capacity to compensate for adversarial noise, amplifying directional shifts in logit space.

After fine-tuning, this correlation weakens significantly ($R^2 = 0.05-0.52$): SwAV achieves only a 17% drop despite retaining $D = 1.7$, implying that fine-tuning enables the head to stabilize predictions even under directional feature

shifts. This pattern aligns with our transfer learning analysis, where fine-tuning disrupted the correlation between ImageNet and downstream instance-specific attack robustness. Regression figures are in Section 2 of the Supp. Mat.

As in Table 2, self-supervised models exhibit higher pre-fine-tuning Magnitude Ratios (e.g., SwAV $M = 17.4$ on ImageNet vs. supervised $M = 3.0$), potentially due to their reliance on globally normalized feature spaces, where FGSM perturbations propagate more aggressively to logits. Fine-tuning reduces M across models (SwAV $M = 2.7$ post-FT), aligning with improved robustness, possibly by suppressing logit magnitude distortions.

These patterns are specific to FGSM, as its single-step gradient reliance may emphasize head-layer instability, whereas iterative attacks could exploit backbone vulnerabilities differently. While our analysis tentatively links directional logit shifts (D) and magnitude sensitivity (M) to robustness, dataset-dependent variations—such as weaker correlations on CIFAR-100 FT—highlight the need for broader evaluation across threat models to generalize these insights.

We also measured inter-class and intra-class distances for both probed and fine-tuned CIFAR-10 representations; however, these metrics did not yield meaningful correlations that explain the observed differences in adversarial robustness. The detailed regression figures, inter/intra-class distance measurements, and t-SNE [82] visualizations of feature spaces are provided in Section 3 of the Supp. Mat.

4.3. Segmentation

Unlike in classification, we observe no strong correlation between ImageNet robustness and segmentation performance across both PASCAL VOC and CityScapes datasets. Self-supervised approaches demonstrate competitive performance in clean conditions, with VICReg leading on PASCAL VOC (77.89 mIOU with frozen backbone) and SwAV on CityScapes (66.48 mIOU). All models suffer catastrophic performance degradation (74-84%) under adversarial attacks regardless of training methodology. This uniform degradation pattern suggests attacks primarily target segmentation modules rather than backbones.

Our experiments with frozen versus unfrozen backbones reveal that frozen backbones generally achieve both higher clean performance and slightly better adversarial robustness on PASCAL VOC. The supervised model’s exception to this pattern stems from using the standard MMSegmentation model for the unfrozen case due to reproduction challenges. These findings indicate that while SSL models produce competitive segmentation performance in clean conditions, they offer minimal advantage in terms of adversarial robustness for segmentation tasks, unlike their significant impact on object recognition. This highlights the need for robustness techniques specifically designed for segmenta-

tion architectures rather than focusing solely on backbone improvements. The full results are in Section 4.2 of the Supp. Mat.

4.4. Detection

Detection results show distinct patterns from recognition and segmentation tasks. On INRIA Person with unfrozen backbones, SwAV demonstrates the highest robustness (72% decrease), while the Supervised model shows the poorest performance (89% decrease). With frozen backbones, Barlow Twins and DINO lead in robustness (both 75% decrease), while VICReg becomes unexpectedly vulnerable (84% decrease), contradicting its recognition performance.

COCO dataset results show improved robustness overall, with Barlow Twins maintaining the strongest performance (50% decrease) and the Supervised model remaining the least robust (67% decrease). This consistent weakness in the Supervised model suggests inherent robustness benefits from self-supervised pretraining.

These findings indicate that task-specific architectures significantly influence adversarial robustness, and robustness in ImageNet classification doesn’t necessarily transfer to detection tasks. This highlights the importance of task-specific evaluations and suggests that backbone architecture becomes less critical than overall model design for downstream applications. The full results are in Section 4.3 of the Supp. Mat.

5. Conclusion

Our evaluation shows that SSL models generally offer greater adversarial robustness than supervised counterparts in image classification, with MoCoV3 performing particularly well—likely due to its momentum encoder and prediction head. This robustness advantage is less evident in segmentation and detection, where task-specific architectures play a larger role. Architectural effects also vary by objective: DINO benefits from ViTs, while MoCoV3 performs better with CNNs. Extended training does not compromise robustness and may slightly improve it. We find that vulnerability to UAPs transfers more consistently across learning paradigms, while susceptibility to IAAs can be mitigated through fine-tuning. These results highlight the complex interplay between SSL objectives, architectures, and downstream tasks, underscoring the need for further study. We hope our findings support continued progress toward more robust visual representation learning.

Acknowledgements

This work was funded by the KUIS AI Center at Koç University, Turkey.

References

- [1] Yuki Markus Asano, Christian Rupprecht, and Andrea Vedaldi. Self-labelling via simultaneous clustering and representation learning, 2020. 2
- [2] Yutong Bai, Jieru Mei, Alan Yuille, and Cihang Xie. Are transformers more robust than cnns?, 2021. 6
- [3] Yang Bai, Yuyuan Zeng, Yong Jiang, Shu-Tao Xia, Xingjun Ma, and Yisen Wang. Improving adversarial robustness via channel-wise activation suppressing, 2022. 1
- [4] Randall Balestriero, Mark Ibrahim, Vlad Sobal, Ari S. Morcos, Shashank Shekhar, Tom Goldstein, Florian Bordes, Adrien Bardes, Grégoire Mialon, Yuandong Tian, Avi Schwarzschild, Andrew Gordon Wilson, Jonas Geiping, Quentin Garrido, Pierre Fernandez, Amir Bar, Hamed Pirsiavash, Yann LeCun, and Micah Goldblum. A cookbook of self-supervised learning. *ArXiv*, abs/2304.12210, 2023. 1
- [5] Adrien Bardes, Jean Ponce, and Yann LeCun. Vircreg: Variance-invariance-covariance regularization for self-supervised learning, 2022. 2, 4
- [6] Yoshua Bengio, Geoffrey Hinton, Andrew Yao, Dawn Song, Pieter Abbeel, Trevor Darrell, Yuval Noah Harari, Ya-Qin Zhang, Lan Xue, Shai Shalev-Shwartz, Gillian Hadfield, Jeff Clune, Tegan Maharaj, Frank Hutter, Atılım Güneş Baydin, Sheila McIlraith, Qiqi Gao, Ashwin Acharya, David Krueger, Anca Dragan, Philip Torr, Stuart Russell, Daniel Kahneman, Jan Brauner, and Sören Mindermann. Managing extreme ai risks amid rapid progress. *Science*, 384(6698): 842–845, 2024. 1
- [7] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kudithipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avani Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models, 2022. 1
- [8] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *Computer Vision – ECCV 2014*, pages 446–461, Cham, 2014. Springer International Publishing. 4
- [9] Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, Percy Liang, and John C. Duchi. Unlabeled data improves adversarial robustness, 2022. 3
- [10] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features, 2019. 2
- [11] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. 2020. 2, 4, 6
- [12] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9630–9640, 2021. 1, 2, 4, 6
- [13] Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay. Adversarial attacks and defences: A survey, 2018. 2
- [14] Ashutosh Chaubey, Nikhil Agrawal, Kavya Barnwal, Keerat K Guliani, and Pramod Mehta. Universal adversarial perturbations: A survey. *arXiv preprint arXiv:2005.08087*, 2020. 2
- [15] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation, 2018. 4
- [16] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. *ArXiv*, abs/2002.05709, 2020. 1, 2, 4, 6
- [17] Tianlong Chen, Sijia Liu, Shiyu Chang, Yu Cheng, Lisa Amini, and Zhangyang Wang. Adversarial robustness: From self-supervised pre-training to fine-tuning, 2020. 3
- [18] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15745–15753, 2020. 2, 5
- [19] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 2, 5
- [20] Xinlei Chen*, Saining Xie*, and Kaiming He. An empirical study of training self-supervised vision transformers. *arXiv preprint arXiv:2104.02057*, 2021. 1, 2, 4, 6
- [21] Prakash Chandra Chhipa, Johan Rodahl Holmgren, Kanjar De, Rajkumar Saini, and Marcus Liwicki. Can self-supervised representation learning methods withstand distribution shifts and corruptions?, 2023. 3
- [22] Mircea Cimpoi, Subhansu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild, 2013. 4

- [23] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding, 2016. 4
- [24] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, pages 886–893 vol. 1, 2005. 4
- [25] Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, Rodolphe Jenatton, Lucas Beyer, Michael Tschanen, Anurag Arnab, Xiao Wang, Carlos Riquelme, Matthias Minderer, Joan Puigcerver, Utku Evci, Manoj Kumar, Sjoerd van Steenkiste, Gamaleldin F. Elsayed, Aravindh Mahendran, Fisher Yu, Avital Oliver, Fantine Huot, Jasmijn Bastings, Mark Patrick Collier, Alexey Gritsenko, Vighnesh Birodkar, Cristina Vasconcelos, Yi Tay, Thomas Mensink, Alexander Kolesnikov, Filip Pavetić, Dustin Tran, Thomas Kipf, Mario Lučić, Xiaohua Zhai, Daniel Keysers, Jeremiah Harmsen, and Neil Houlsby. Scaling vision transformers to 22 billion parameters, 2023. 6
- [26] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. 2
- [27] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum, 2018. 4
- [28] 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, 2021. 1, 4, 6
- [29] Linus Ericsson, Henry Gouk, and Timothy M. Hospedales. How well do self-supervised models transfer?, 2021. 4
- [30] Aleksandr Ermolov, Aliaksandr Siarohin, Enver Sanginetto, and Nicu Sebe. Whitening for self-supervised representation learning, 2021. 2
- [31] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. <http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html>. 4
- [32] Alhussein Fawzi, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. Robustness of classifiers: from adversarial to random noise, 2016. 1
- [33] Li Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 Conference on Computer Vision and Pattern Recognition Workshop*, pages 178–178, 2004. 4
- [34] Micah Goldblum, Hossein Souri, Renkun Ni, Manli Shu, Viraj Prabhu, Gowthami Somapalli, 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, 2023. 3
- [35] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 1, 4
- [36] Jean-Bastien Grill, Florian Strub, Florent Althé, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. *ArXiv, abs/2006.07733*, 2020. 2, 4
- [37] Rohit Gupta, Naveed Akhtar, Ajmal Mian, and Mubarak Shah. Contrastive self-supervised learning leads to higher adversarial susceptibility, 2022. 2, 4
- [38] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 1, 4
- [39] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722*, 2019. 2, 5
- [40] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. Masked autoencoders are scalable vision learners. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15979–15988, 2021. 1, 2, 6
- [41] Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty, 2019. 3
- [42] Chih-Hui Ho and Nuno Vasconcelos. Contrastive learning with adversarial examples, 2020. 2, 4
- [43] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. 1
- [44] Hao Huang, Ziyang Chen, Huanran Chen, Yongtao Wang, and Kevin Zhang. T-sea: Transfer-based self-ensemble attack on object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20514–20523, 2023. 4
- [45] Minyoung Huh, Pulkit Agrawal, and Alexei A. Efros. What makes imagenet good for transfer learning?, 2016. 4
- [46] Mark Ibrahim, David Klindt, and Randall Balestriero. Occam’s razor for self supervised learning: What is sufficient to learn good representations?, 2024. 5
- [47] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features, 2019. 1
- [48] Ziyang Jiang, Tianlong Chen, Ting Chen, and Zhangyang Wang. Robust pre-training by adversarial contrastive learning, 2020. 2, 4
- [49] Minseon Kim, Jihoon Tack, and Sung Ju Hwang. Adversarial self-supervised contrastive learning, 2020. 2, 4
- [50] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, Quan Vuong, Thomas Koliar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. Openvla: An open-source vision-language-action model, 2024. 1
- [51] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. 4

- [52] Antoni Kowalczyk, Jan Dubiński, Atiyeh Ashari Ghomi, Yi Sui, George Stein, Jiapeng Wu, Jesse C. Cresswell, Franziska Boenisch, and Adam Dziedzic. Benchmarking robust self-supervised learning across diverse downstream tasks, 2024. 1, 3
- [53] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *2013 IEEE International Conference on Computer Vision Workshops*, pages 554–561, 2013. 4
- [54] Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009. 4
- [55] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 6
- [56] Ang Li, Yifei Wang, Yiwen Guo, and Yisen Wang. Adversarial examples are not real features, 2024. 1
- [57] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. 4
- [58] Rundong Luo, Yifei Wang, and Yisen Wang. Rethinking the effect of data augmentation in adversarial contrastive learning, 2023. 3
- [59] 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. 4
- [60] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft, 2013. 4
- [61] Warren Morningstar, Alex Bijamov, Chris Duvarney, Luke Friedman, Neha Kalibhat, Luyang Liu, Philip Mansfield, Renan Rojas-Gomez, Karan Singhal, Bradley Green, and Sushant Prakash. Augmentations vs algorithms: What works in self-supervised learning, 2024. 5
- [62] Muzammal Naseer, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Fatih Porikli. A self-supervised approach for adversarial robustness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 262–271, 2020. 6
- [63] Yurii Nesterov. A method for solving the convex programming problem with convergence rate $o(1/k^2)$. *Proceedings of the USSR Academy of Sciences*, 269:543–547, 1983. 4
- [64] A. Tuan Nguyen, Ser Nam Lim, and Philip Torr. Task-agnostic robust representation learning, 2022. 2, 4
- [65] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. 1, 2, 6
- [66] Utku Ozbulak, Hyun Jung Lee, Beril Boga, Esla Timothy Anzaku, Homin Park, Arnout Van Messem, Wesley De Neve, and Joris Vankerschaver. Know your self-supervised learning: A survey on image-based generative and discriminative training, 2023. 4
- [67] Serdar Ozsoy, Shadi Hamdan, Sercan Ö. Arik, Deniz Yuret, and Alper T. Erdogan. Self-supervised learning with an information maximization criterion, 2022. 2
- [68] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library, 2019. 4
- [69] Francesco Pinto, Philip H. S. Torr, and Puneet K. Dokania. An impartial take to the cnn vs transformer robustness contest, 2022. 6
- [70] Maura Pintor, Fabio Roli, Wieland Brendel, and Battista Biggio. Fast minimum-norm adversarial attacks through adaptive norm constraints, 2021. 4
- [71] Alec Radford and Karthik Narasimhan. Improving language understanding by generative pre-training. 2018. 2
- [72] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks, 2016. 4
- [73] Jérôme Rony, Jean-Christophe Pesquet, and Ismail Ben Ayed. Proximal splitting adversarial attacks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 4
- [74] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge, 2015. 2, 3, 4
- [75] Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. Adversarially robust generalization requires more data, 2018. 1
- [76] Ali Shafahi, W. Ronny Huang, Christoph Studer, Soheil Feizi, and Tom Goldstein. Are adversarial examples inevitable?, 2020. 1
- [77] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *Neural Information Processing Systems*, 2016. 5
- [78] 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. 1
- [79] Thomas Tanay and Lewis Griffin. A boundary tilting perspective on the phenomenon of adversarial examples, 2016. 1
- [80] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023. 1

- [81] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019. [5](#)
- [82] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9 (86):2579–2605, 2008. [8](#)
- [83] Yisen Wang, Difan Zou, Jinfeng Yi, James Bailey, Xingjun Ma, and Quanquan Gu. Improving adversarial robustness requires revisiting misclassified examples. In *ICLR*, 2020. [1](#)
- [84] Yisen Wang, Xingjun Ma, James Bailey, Jinfeng Yi, Bowen Zhou, and Quanquan Gu. On the convergence and robustness of adversarial training, 2022.
- [85] Dongxian Wu, Shu tao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization, 2020. [1](#)
- [86] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan Yuille. Adversarial examples for semantic segmentation and object detection, 2017. [4](#)
- [87] Cihang Xie, Mingxing Tan, Boqing Gong, Jiang Wang, Alan Yuille, and Quoc V. Le. Adversarial examples improve image recognition, 2020. [2, 4](#)
- [88] Xilie Xu, Jingfeng Zhang, Feng Liu, Masashi Sugiyama, and Mohan Kankanhalli. Efficient adversarial contrastive learning via robustness-aware coreset selection, 2023. [3](#)
- [89] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction, 2021. [2, 4](#)
- [90] Chaoning Zhang, Kang Zhang, Chenshuang Zhang, Axi Niu, Jiu Feng, Chang D. Yoo, and In So Kweon. Decoupled adversarial contrastive learning for self-supervised adversarial robustness, 2022. [2, 4](#)
- [91] Yuanyi Zhong, Haoran Tang, Junkun Chen, Jian Peng, and Yu-Xiong Wang. Is self-supervised learning more robust than supervised learning?, 2022. [3](#)