Evaluating the Adversarial Robustness of Image - Joint Embedding Predictive Architectures

Aditya Agarwala
MTech AI
Indian Institute of Science
Bengaluru, KA, 560012
adityaa1@iisc.ac.in

Purva Parmar
MTech AI
Indian Institute of Science
Bengaluru, KA, 560012
purvaparmar@iisc.ac.in

Abstract

Modern vision models have achieved impressive performance yet remain highly susceptible to adversarial perturbations, which raises concerns about their deployment in safety-critical settings. Recent self-supervised learning methods have been hypothesized to improve robustness by emphasizing semantic structure rather than low-level pixel statistics. In this work, we investigate whether the Image Joint-Embedding Predictive Architecture (I-JEPA), a non-contrastive predictive self-supervised method, provides greater adversarial robustness compared to standard supervised pretraining. Using a controlled linear probing setup on ViT-H/14 backbones and two downstream classification tasks, CIFAR-100 and ImageNet-100, we evaluate robustness under a suite of gradient-based and ensemble attacks. Our results show a mixed robustness profile. I-JEPA models consistently outperform supervised counterparts under weaker, single-step attacks and under multi-step attacks when operating on larger and semantically richer images. However, these benefits diminish or reverse on small-resolution datasets such as CIFAR-100. These findings suggest that any robustness advantages of I-JEPA representations likely depend on input scale and structural complexity, which motivates further investigation into the conditions under which predictive self-supervised objectives promote adversarial resilience.

1 Introduction

Deep learning has transformed computer vision, driving advances in domains such as medical imaging, autonomous navigation, and industrial inspection (Zhou et al., 2020; Hütten et al., 2024; Grigorescu et al., 2019). However, despite their impressive performance, modern vision models remain highly vulnerable to adversarial attacks: small, carefully crafted perturbations to input images that can cause misclassifications while remaining visually imperceptible to humans (Szegedy et al., 2014; Goodfellow et al., 2015). This fragility raises serious concerns about the reliability and safety of deploying such systems in real-world, safety-critical settings.

Recent progress in self-supervised learning (SSL) has sparked interest in whether learning from data structure rather than explicit labels can yield more robust and semantically meaningful representations. SSL methods encourage models to capture high-level abstractions of the visual world, potentially reducing overreliance on low-level statistical cues often exploited by adversarial perturbations (Chen et al., 2020a; Mohseni et al., 2020).

Among SSL approaches, the Image Joint-Embedding Predictive Architecture (I-JEPA) (Assran et al., 2023) introduces a distinct paradigm. Instead of reconstructing pixels or predicting augmented views, I-JEPA learns by predicting the latent representations of masked regions from their surrounding context. This objective shifts the learning focus from local texture statistics to global semantic understanding. As a result, we hypothesize that I-JEPA-trained

models may possess intrinsic robustness, developing representations that are less sensitive to adversarial perturbations at the pixel level.

Despite this theoretical motivation, the robustness of I-JEPA models has not been systematically validated. Existing evaluations have largely focused on accuracy and transfer performance, leaving open the question of whether the predictive objective indeed provides greater adversarial resilience compared to standard supervised pretraining.

To address this gap, our study performs a systematic comparison between Vision Transformers (ViT) (Dosovitskiy et al., 2021) pretrained with the I-JEPA objective and their supervised counterparts of identical architecture. By training linear classification heads and evaluating their performance under established adversarial attack methods, we aim to empirically assess whether the hypothesized robustness advantage of I-JEPA manifests in practice.

We release the code in a GitHub repository: `https://github.com/TheReconPilot/ijepa-adversarial-robustness`.

2 Related Work

2.1 Adversarial Robustness in Supervised Models

Deep neural networks are highly vulnerable to adversarial examples like small, imperceptible perturbations that can cause misclassification. The Fast Gradient Sign Method (FGSM) (Goodfellow et al., 2015) is an example of a simple one-step gradient-based attack, later extended by Projected Gradient Descent (PGD) (Madry et al., 2019), which iteratively applies perturbations while constraining them within a bounded ℓ_p-norm ball. PGD has long served as the standard benchmark for adversarial robustness and forms the basis of adversarial training techniques that trade clean accuracy for robustness (Zhang et al., 2019; Tsipras et al., 2019).

Beyond these basic first-order attacks, a large body of work has explored stronger and more diverse threat models. Optimization-based attacks such as the Carlini&Wagner attack (Carlini & Wagner, 2017) and DeepFool (Moosavi-Dezfooli et al., 2016) explicitly minimize perturbation magnitude subject to misclassification, often producing smaller and more transferable perturbations. Decision-based and score-based black-box attacks, including the Boundary Attack (Brendel et al., 2018) and Natural Evolutionary Strategies (NES)-type gradient estimation methods (Ilyas et al., 2018), relax the assumption of full gradient access and instead rely on model outputs or labels only, expanding robustness evaluations to settings where internal gradients are unavailable. More recent work has focused on improving attack efficiency and reliability under adaptive defenses, for instance via loss formulations that reduce gradient obfuscation (Uesato et al., 2018; Tramèr et al., 2020) or more robust optimization objectives such as the Difference of Logits Ratio (DLR) loss (Croce & Hein, 2020b).

In parallel with the development of individual attacks, there has been growing recognition that reliable robustness evaluation requires ensembles of strong, complementary attacks and careful hyperparameter tuning (Carlini et al., 2019; Croce & Hein, 2020b). AutoAttack (Croce & Hein, 2020b) embodies this philosophy by providing a parameter-free benchmark suite composed of several state-of-the-art attacks: Auto-PGD with cross-entropy (APGD-CE) and DLR (APGD-DLR) losses for white-box ℓ_p-constrained perturbations, the Fast Adaptive Boundary (FAB) attack for finding minimal-norm adversarial examples close to decision boundaries (Croce & Hein, 2020a), and the Square attack as a score-based black-box method (Andriushchenko et al., 2020). AutoAttack is now widely adopted as a standardized robustness benchmark, serving as the default evaluation protocol in robustness leaderboards and being treated as a reference method in recent benchmarks, surveys, and defense papers (Croce et al., 2020; Gao et al., 2022; Lorenz et al., 2021; Piras et al., 2025). We use AutoAttack as the primary evaluation attack to obtain a strong and reproducible robustness baseline in our experiments.

2.2 Adversarial Robustness in Self-Supervised Learning

SSL learns transferable representations from unlabeled data via pretext tasks such as rotation prediction (Gidaris et al., 2018), contrastive learning (Chen et al., 2020b; He et al., 2020), clustering-based objectives (Caron et al., 2018; 2020), and masked reconstruction or joint-embedding predictive architectures (He et al., 2022; Grill et al., 2020; Assran et al., 2023). When pre-trained on large-scale datasets like CIFAR-10/CIFAR-100 (Krizhevsky et al., 2009) and ImageNet (Deng et al., 2009), these methods often rival or surpass supervised pre-training on downstream tasks such as ImageNet supervised and semi-supervised classification, as well as transfer to object detection and segmentation benchmarks (Chen et al., 2020b; Grill et al., 2020).

Early evidence that SSL can benefit robustness came from adding self-supervised auxiliary tasks to standard supervised training. When rotation prediction is combined with supervised learning on CIFAR-10 and ImageNet, improvements are observed in both corruption robustness and adversarial robustness, as measured under ℓ_∞-bounded FGSM/PGD attacks and common corruption benchmarks (Hendrycks et al., 2019). Related approaches leverage self-supervision to improve out-of-distribution detection (Mohseni et al., 2020) and to exploit unlabeled data for robust training (Carmon et al., 2019). Other works integrate adversarial perturbations directly into the SSL objective: adversarial contrastive pre-training (Jiang et al., 2020) and adversarially robust SSL methods (Chen et al., 2020a; Kim et al., 2020) apply PGD-style perturbations during pre-training and report consistent gains in robust accuracy under ℓ_∞ PGD and, in some cases, AutoAttack (Madry et al., 2019; Croce & Hein, 2020b) on CIFAR-10/100. For masked autoencoders (MAE), robustness on ImageNet has also been improved by a test-time procedure that slightly modifies inputs in the frequency domain ("frequency-domain prompting"), making MAE-style models more resistant to gradient-based adversarial attacks (He et al., 2022; Huang et al., 2023). Complementary studies evaluate SSL representations under distribution shifts and corruptions, generally finding that self-supervised features can yield more stable performance than purely supervised baselines (Chhipa et al., 2023; Gui et al., 2024).

Recent work examines adversarial robustness of discriminative SSL encoders in a more systematic way. Contrastive and clustering-based SSL methods (e.g., SimCLR, MoCo, SwAV, DeepCluster) are compared to supervised baselines on CIFAR-10/CIFAR-100 and ImageNet in Çağatan et al. (2025), and robust accuracy is measured under strong ℓ_∞ attacks such as multi-step PGD and AutoAttack (Madry et al., 2019; Croce & Hein, 2020b). They find that SSL-pretrained models often exhibit stronger robustness in the *linear-probe* setting, where a frozen backbone is evaluated via a trained linear classifier, suggesting that the learned representations encode invariances that are beneficial for adversarial robustness. However, these gains largely diminish and sometimes disappear or reverse after end-to-end fine-tuning, indicating that the supervised fine-tuning objective can overwrite or weaken the robustness properties of the pretrained representation. This observation is consistent with results showing that contrastive SSL can, in some regimes, lead to higher adversarial susceptibility after fine-tuning (Gupta et al., 2023).

2.3 I-JEPA and Adversarial Robustness

Most existing empirical robustness studies focus on contrastive, clustering, or MAE-style SSL models, building on contrastive frameworks such as SimCLR (Chen et al., 2020b), clustering-based methods such as SwAV (Caron et al., 2020), and masked autoencoders (MAE) (He et al., 2022). In contrast, recent joint-embedding predictive architectures such as I-JEPA (Assran et al., 2023) remain largely unexplored from a robustness perspective. I-JEPA differs from these methods by predicting latent representations of masked image regions from visible context in a joint-embedding space, without contrastive negatives or pixel-level reconstruction, and achieves competitive or state-of-the-art transfer performance on ImageNet with ViT backbones (Assran et al., 2023; Dosovitskiy et al., 2021). Because I-JEPA is trained with a very different objective than contrastive or reconstruction-based SSL, it is not clear a priori whether the features it learns should be more or less robust to adversarial perturbations.

From an applications standpoint, I-JEPA style encoders are being positioned as general-purpose vision backbones, similar to contrastive and masked-autoencoding models. Existing surveys of SSL and robustness mention I-JEPA style models only briefly and do not provide a dedicated robustness analysis (Gui et al., 2024); to our knowledge, the only public robustness discussion of I-JEPA is an informal, non-peer-reviewed case study (Kulm, 2023).

This combination of (i) a qualitatively different pretext objective, (ii) growing practical interest, and (iii) a lack of systematic empirical evidence makes I-JEPA an important test case for understanding adversarial robustness in modern SSL. In this work, we therefore focus specifically on the adversarial robustness of I-JEPA-style encoders under strong white and black box attacks, using AutoAttack as our primary evaluation attack to obtain a reliable robustness baseline.

3 Methods and Experimental Setup

3.1 Linear Probing Setup

We evaluate the representational robustness of pretrained Vision Transformers by adopting a linear probing setup, which isolates the quality of learned features from the effects of fine-tuning. In linear probing, for each model, all transformer layers are frozen and a single linear classification head is added and trained on downstream datasets. Specifically, from HuggingFace, we use two publicly available ViT-H/14 backbones of identical architecture, both pretrained on ImageNet-21k (Deng et al., 2009), but with differing pretraining objectives:

- Supervised model ("ViT"): `google/vit-huge-patch14-224-in21k`, using the cross-entropy objective for classification
- I-JEPA model ("I-JEPA"): `facebook/ijepa_vith14_22k`, using the Image Joint-Embedding Predictive Architecture (I-JEPA) objective.

We use two datasets of varying scale and granularity: **CIFAR-100** (Krizhevsky et al., 2009) and **`ilee0022/ImageNet100`** from HuggingFace, a subset of ImageNet-1k (Russakovsky et al., 2015), whose test set does not overlap with the ImageNet-21k (Deng et al., 2009) pretraining data. Both datasets are split into training and test sets following standard practice.

We adapt the linear probing architecture for each model following the evaluation protocols established in the original I-JEPA paper (Assran et al., 2023), described as follows. As the models process global context differently, we employ the specific feature extraction strategy best suited for each architecture:

- **Supervised ViT**: As this model is trained with a dedicated class token, we extract the [CLS] embedding from the final transformer layer and apply a one-dimensional Batch Normalization layer (Ioffe & Szegedy, 2015) to the features before passing them to the linear classifier. We show the comparison of this ViT model with and without the Batch Norm in Appendix A.
- **I-JEPA**: Since I-JEPA does not utilize a [CLS] token, we compute the global average pooling of the patch embeddings of the last layer patches, or alternatively the concatenation of average pooled last four layer patches. We show the comparison of both of these strategies on classification performance in Appendix A.

Training is performed using the cross-entropy loss, optimized with AdamW optimizer (Loshchilov & Hutter, 2019). We also use PyTorch's Automatic Mixed Precision while training. Input images are resized to 224×224 with Bicubic interpolation. The classification head is trained for 25 epochs with checkpoints saved at every epoch. The learning rate is cosine-decayed with warmup. We pick the best performing checkpoint. The resulting models are used for subsequent adversarial robustness evaluations.

3.2 Adversarial Evaluation

To assess model robustness, we generate adversarial examples using a suite of gradient-based and black-box attack methods, ranging from single-step perturbations to iterative ensembles.

3.2.1 Standard Gradient Attacks

Fast Gradient Sign Method (FGSM) (Goodfellow et al., 2015) is a single-step attack that efficiently perturbs the input image x in the direction of the loss gradient:

$$x_{\text{adv}} = \text{Clip}_{[0,1]}\left(x + \epsilon \cdot \text{sign}(\nabla_x \mathcal{L}(x, y))\right), \tag{1}$$

where y is the true label, $\mathcal{L}$ is the cross-entropy loss, and ϵ bounds the perturbation magnitude in the ℓ_∞ norm. We also clip the resulting image to ensure pixel values remain within $[0, 1]$ and obtain x_{adv}. We evaluate FGSM at perturbation budgets $\epsilon \in \{1/255, 2/255, 4/255, 8/255\}$.

Projected Gradient Descent (PGD) (Madry et al., 2019) takes this further by applying FGSM iteratively. In each step t, the image is updated and projected back onto the allowable perturbation set $\mathcal{B}_\epsilon(x)$:

$$x_{t+1} = \Pi_{\mathcal{B}_\epsilon(x)}\left(x_t + \alpha \cdot \delta_t\right), \tag{2}$$

where α is the step size. We evaluate PGD under two distinct norm constraints to capture different aspects of perceptual robustness:

- **ℓ_∞-PGD:** The perturbation is bounded by the maximum pixel change ($\max_i |x_i - x_i'| \leq \epsilon$). Here, $\delta_t = \text{sign}(\nabla_{x_t} \mathcal{L})$, allowing uniform noise across the image.
- **ℓ_2-PGD:** The perturbation is bounded by Euclidean distance ($\|x - x'\|_2 \leq \epsilon$). Here, $\delta_t = \frac{\nabla_{x_t} \mathcal{L}}{\|\nabla_{x_t} \mathcal{L}\|_2}$, forcing the attack to concentrate changes on pixels with the highest gradient sensitivity rather than spreading them uniformly. This can be considered a proxy for more naturalistic energy-bounded distortions.

For PGD, we perform attacks with step counts of 10, using 3 random restarts because the attack landscape is non-convex and single-start PGD often misses adversarial examples (Madry et al., 2019). It also helps against a phenomenon known as gradient masking (Athalye et al., 2018) where a model "hides" its vulnerability to adversarial attacks by making its gradients useless for optimization, even though the model itself remains vulnerable.

3.2.2 AutoAttack Ensemble

AutoAttack (Croce & Hein, 2020b), a hyperparameter-free ensemble of four diverse attacks. The work introduces two methods: APGD-CE and APGD-DLR as described below, and combines two more existing methods of FAB and Square Attacks. AutoAttack executes these methods sequentially and only passes the survivors to the next attack. The ensemble consists of:

1. **APGD-CE (Auto-PGD Cross-Entropy):** A variation of PGD that automatically adjusts the step size α over time to escape local maxima, removing the need for manual tuning. It maximizes the standard cross-entropy loss.
2. **APGD-DLR (Auto-PGD Difference of Logits Ratio):** This variant uses the Difference of Logits Ratio (DLR) loss, which is scale-invariant and specifically designed to prevent gradient vanishing when the model predicts the correct class with high confidence. The loss is defined as:

$$\text{DLR}(x, y) = -\frac{z_y - \max_{i \neq y} z_i}{z_{\pi_1} - z_{\pi_3}}, \tag{3}$$

where z are the logits and π is the sorting permutation of components. This component targets "hard" samples where standard cross-entropy fails.

3. **FAB (Fast Adaptive Boundary):** The FAB attack (Croce & Hein, 2020a) is a geometric approach that aims to find the minimal perturbation necessary to cross the decision boundary, rather than maximizing loss within a fixed ϵ-ball.
4. **Square Attack:** A score-based black-box attack (Andriushchenko et al., 2020) that does not rely on gradient information. It uses random search to update localized square-shaped patches of the image. This component ensures the evaluation remains valid even if the model gradients are obfuscated or non-useful (gradient masking).

We take three models which have been trained with a linear probe: the supervised ViT model (Google ViT) with batch normalization applied before the classification head, the I-JEPA model with average pooling of last layer (I-JEPA Last-1) and the I-JEPA model with the concatenation of last four average pooled layers (I-JEPA Last-4). We subject these to FGSM, PGD with ℓ_∞ and ℓ_2 norms, APGD-CE, APGD-DLR, and AutoAttack (as an ensemble). All of the attacks are performed in Full Precision (FP32).

We report the robust accuracy as the percentage of images that remain correctly classified after the execution of each attack. Due to computational constraints and some attacks (like FAB and Square) being really computationally intensive, we have been unable to report robust accuracies for all configurations.

The experiments were run on NVIDIA RTX A5000 and NVIDIA Tesla V100 GPUs on servers, courtesy of the MTech AI Office of the Indian Institute of Science.

4 Results

We report the results for three variants in Table 1. The Google ViT is the standard ViT-H/14 model, but with a batch norm added before training a classfication head. The I-JEPA Last-1 is the I-JEPA ViT model with average pooling of last layer patches before the classification head. The I-JEPA Last-4 is the I-JEPA ViT model with concatenation of average-pooled last four layers. The results are reported on CIFAR-100 and ImageNet-100.

	CIFAR-100			ImageNet-100		
Attack	**Google ViT**	**I-JEPA Last-4**	**I-JEPA Last-1**	**Google ViT**	**I-JEPA Last-4**	**I-JEPA Last-1**
Clean	93.68 (-)	85.42 (-)	76.94 (-)	89.78 (-)	88.46 (-)	85.79 (-)
FGSM ($\epsilon = 1/255$)	13.68 (85.4)	25.32 (70.4)	24.56 (68.1)	39.14 (56.4)	55.25 (37.5)	57.38 (33.1)
FGSM ($\epsilon = 2/255$)	8.16 (91.3)	19.92 (76.7)	19.92 (74.1)	26.62 (70.4)	40.83 (53.8)	46.01 (46.4)
FGSM ($\epsilon = 4/255$)	7.10 (92.4)	19.94 (76.7)	20.04 (74.0)	19.33 (78.5)	30.73 (65.3)	37.81 (55.9)
FGSM ($\epsilon = 8/255$)	6.68 (92.9)	19.00 (77.8)	15.56 (79.8)	16.09 (82.1)	26.68 (69.8)	33.08 (61.4)
PGD L_2 ($\epsilon = 0.5$)	5.90 (93.7)	2.30 (97.3)	2.74 (96.4)	28.82 (67.9)	54.64 (38.2)	55.61 (35.2)
PGD L_2 ($\epsilon = 1.0$)	-	-	-	4.63 (94.8)	26.75 (69.8)	31.76 (63.0)
PGD L_∞ ($\epsilon = 0.2/255$)	44.80 (52.2)	32.18 (62.3)	27.24 (64.6)	70.45 (21.5)	71.85 (18.8)	73.67 (14.1)
PGD L_∞ ($\epsilon = 1/255$)	1.22 (98.7)	0.66 (99.2)	0.60 (99.2)	11.69 (87.0)	30.02 (66.1)	31.86 (62.9)
PGD L_∞ ($\epsilon = 2/255$)	0.00 (100)	0.08 (99.9)	0.06 (99.9)	0.51 (99.4)	7.24 (91.8)	10.03 (88.3)
PGD L_∞ ($\epsilon = 8/255$)	0.00 (100)	0.00 (100)	0.00 (100)	0.00 (100)	0.70 (99.2)	0.15 (99.8)
APGD-CE	0.20 (99.8)	0.04 (100)	-	2.99 (96.7)	15.57 (82.4)	15.39 (82.1)
APGD-DLR	0.30 (99.7)	0.00 (100)	-	4.54 (94.9)	18.77 (78.8)	19.91 (76.8)
AutoAttack	0.18 (99.8)	0.00 (100)	-	-	-	-

Table 1: Adversarial Attack Results for CIFAR-100 and ImageNet-100. Values are reported as: **Top-1 Accuracy (Drop %)**. For the Top-1 Accuracy, higher is better, and for the Drop, lower is better. All PGD attacks are performed with 10 steps and 3 random restarts. All FGSM and AutoAttack related attacks are performed on L_∞ norm.

A few major differences from the results reported in the Mid-term Report are that the Google ViT did not have Batch Norm, and the I-JEPA ViT models did not have any kind of average pooling. Here, more attacks and configurations have been added, and all attacks have been performed in full precision (fp32), which additionally improves the clean performance too. The clean performance when evaluated on automatic mixed precision can be found in Appendix A.

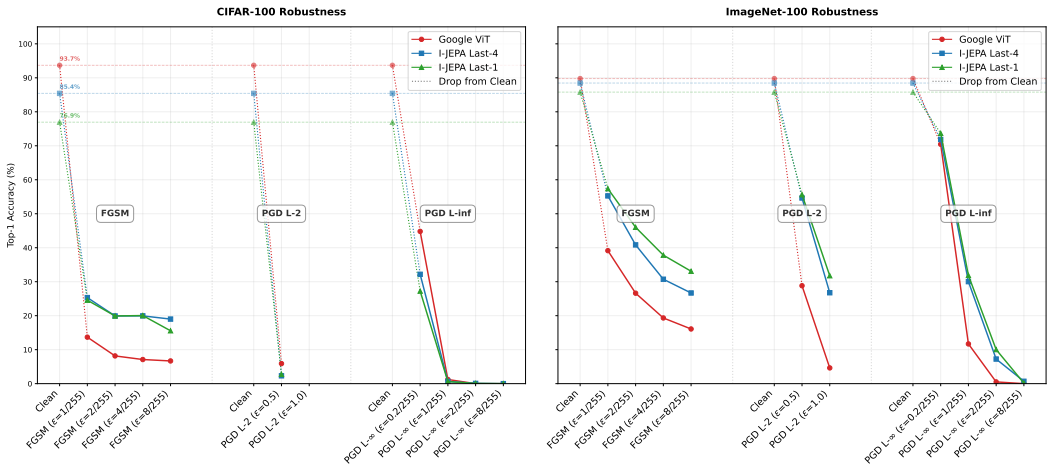

Figure 1: Classification Accuracy of the different models against different attacks. Clean represents the baseline model performance without any attack.

In terms of adversarial robustness, the I-JEPA models consistently fare better than the standard Google ViT model on a single-step gradient based attack like FGSM.

When subjected to a strong multi-step attack like PGD on CIFAR-100, all the models drastically drop in performance. The value of ϵ specifies how big of a perturbation is allowed. The drop is still bigger for I-JEPA models on CIFAR-100. A possible reason for this is how I-JEPA models handle images. I-JEPA learns to predict missing patches of images from the surrounding context, all in a latent space. The CIFAR-100 images are 32×32 in size, and need to be resized to 224×224 using a Bicubic Interpolation and a lot of the same features are repeated in a block-like fashion. The I-JEPA latent representation likely becomes texture dependent and fragile as the images themselves have very little local detail.

For ImageNet-100, the I-JEPA models still fare consistently better than the Google ViT model on all attacks, with a much lesser relative drop in performance. This may be due to the I-JEPA latent representations now being able to learn coarser and more semantic features due to larger images. This likely results in the I-JEPA models being more resistant to local pixel level noise.

Between the Last-1 and Last-4 variants of I-JEPA, the Last-1 variant marginally performs better too, showing smaller drops, even though it starts with a relatively lower clean accuracy.

For PGD, the L_2 norm attacks result in more object-level distortions than pixel-level noise. I-JEPA's features are learnt in a latent space, and they seem to resist the attacks better than a standard ViT trained on a supervised objective which has learnt features in the final image/pixel-space.

Overall, I-JEPA's adversarial robustness benefits seem to manifest when images have sufficient global structure (i.e. larger and more detailed images), but still break down when the images are small enough.

5 Limitations and Future Work

Our study faces limitations stemming primarily from computational constraints and the scope of the current investigation. First, although AutoAttack comprises a diverse ensemble of attacks, including APGD-CE, APGD-DLR, FAB, and Square, we were unable to report full results for the latter two. FAB and Square attacks are significantly more computationally demanding, and executing them at scale proved infeasible within the available time and resources. Consequently, we report detailed results only for the APGD variants and the

overall AutoAttack ensemble, which itself mitigates cost by sequentially passing only the surviving adversarial examples to subsequent attacks. Even with these measures, generating complete results for a dataset as large as ImageNet-100, coupled with models not optimized for efficient inference, remained computationally intractable.

Second, while AutoAttack includes both white-box and black-box components, our reported results primarily reflect white-box settings, which assume complete access to model parameters. The limited inclusion of black-box attacks restricts the breadth of robustness conclusions and represents an avenue for future extension.

Third, our analysis is conducted exclusively under a linear probing setup, in which all backbone parameters are frozen and only a classification head is trained. Although this isolates representational robustness, it does not account for how these models might behave under full or partial fine-tuning. Exploring fine-tuning–based robustness, especially adversarial fine-tuning, was beyond the scope of this project given resource limitations, but constitutes an important direction for follow-up work.

Finally, our evaluation focuses on two specific Vision Transformers: a Google-provided ViT-H/14 and a Meta-provided I-JEPA ViT-H/14, both pretrained on ImageNet-21k. Broader comparisons with other self-supervised or foundation models, such as MAE, DINOv2, SimCLR and MoCo were excluded for feasibility reasons but remain an important extension for future study.

6 Conclusion

This study provided an initial systematic evaluation of the adversarial robustness of I-JEPA pretrained ViT compared to a supervised vanilla ViT model. Across a range of attacks and datasets, we observed that the robustness benefits of I-JEPA were not universal but instead depended on the characteristics of the input data and the strength of the adversarial attack.

On CIFAR-100, I-JEPA models had lower clean accuracy and often exhibited sharper relative performance drops under stronger attacks such as multi-step PGD and AutoAttack. The visual analysis in Figure 1 suggested that the required upscaling of 32×32 images to 224×224 introduced repetitive texture patterns that interacted poorly with the I-JEPA predictive objective. As a result, the learned latent representations appeared fragile in settings where images provided limited local detail.

On ImageNet-100, where images contained richer global structure, I-JEPA variants consistently retained higher robust accuracy across FGSM, PGD under both norms, and APGD attacks. These results indicate that predictive self-supervision encouraged the learning of more semantically grounded features that were less sensitive to local pixel-level perturbations when the images offered sufficient spatial complexity.

Overall, I-JEPA did not yield uniformly superior adversarial robustness. However, it demonstrated promising robustness characteristics in settings with larger and more detailed images and under specific perturbation regimes. This study highlighted that robustness in predictive self-supervised learning was nuanced and shaped by dataset scale, representational granularity, and the nature of adversarial attacks. Future work should extend these evaluations to full fine-tuning, additional SSL models, and broader attack suites to obtain a more complete understanding of the robustness properties of predictive self-supervised models.

Acknowledgements

We are grateful to Prof. Danish Pruthi and the Anudeep, the Teaching Assistant, for the wonderful DS 307: Ethics in AI course at the Indian Institute of Science, Bengaluru. We are also grateful for their insightful feedback during the early stages of this work. We also thank all the course participants for the engaging discussions that enriched our understanding of the subject.

Finally, we acknowledge the support of the MTech AI Office for providing GPU-enabled servers. This computational infrastructure was essential for carrying out our study.

Author Contributions

Purva Parmar and Aditya Agarwala are equal contributors in preparing this report.

Declaration about AI Assistance

We acknowledge the use of AI assistants (such as OpenAI's ChatGPT, Google's Gemini, xAI's Grok, and Anthropic's Claude) in this project. The AI's role involved assisting with the experimental setup through code generation and editing, as well as helping to refine the initial draft of this report for clarity and conciseness. All AI-generated contributions were supervised, reviewed, and adapted by the human authors, who are fully responsible for the final report.

References

Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: a query-efficient black-box adversarial attack via random search. In *European conference on computer vision*, pp. 484–501. Springer, 2020.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture, 2023. URL https://arxiv.org/abs/2301.08243.

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International conference on machine learning*, pp. 274–283. PMLR, 2018.

Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models, 2018. URL https://arxiv.org/abs/1712.04248.

Ömer Veysel Çağatan, Ömer Faruk Tal, and M Emre Gürsoy. Adversarial robustness of discriminative self-supervised learning in vision. *arXiv preprint arXiv:2503.06361*, 2025.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks, 2017. URL https://arxiv.org/abs/1608.04644.

Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, Aleksander Madry, and Alexey Kurakin. On evaluating adversarial robustness, 2019. URL https://arxiv.org/abs/1902.06705.

Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, Percy Liang, and John C. Duchi. Unlabeled data improves adversarial robustness. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/1ef91c212e30e14bf125e9374262401f-Abstract.html.

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 132–149, 2018.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 9912–9924, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/75eb86ddbc1cc57f15e1d4f3c4568bfc-Abstract.html.

Tianlong Chen, Sijia Liu, Shiyu Chang, Yu Cheng, Lisa Amini, and Zhangyang Wang. Adversarial robustness: From self-supervised pre-training to fine-tuning, 2020a. URL https://arxiv.org/abs/2003.12862.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020b. URL https://arxiv.org/abs/2002.05709.

Pratik Chandra Chhipa, Sivalogeswaran Sankaran, Ivica Purohit, Madhusudan Narotam, Rakesh Kapoor, Aditya Seshadri, Suyash Gupta, Debdoot Ghosh, Mehul Mendiratta, and Biplab Banerjee. Can self-supervised representation learning methods withstand distribution shifts and corruptions? In *ICCV 2023 Workshop on Out-of-Distribution Generalization (OOD-CV)*, 2023.

Francesco Croce and Matthias Hein. Minimally distorted adversarial examples with a fast adaptive boundary attack. In *International conference on machine learning*, pp. 2196–2205. PMLR, 2020a.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pp. 2206–2216. PMLR, 2020b.

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

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, 2009. doi: 10.1109/CVPR.2009.5206848.

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. URL https://arxiv.org/abs/2010.11929.

Ruize Gao, Jiongxiao Wang, Kaiwen Zhou, Feng Liu, Binghui Xie, Gang Niu, Bo Han, and James Cheng. Fast and reliable evaluation of adversarial robustness with minimum-margin attack, 2022. URL https://arxiv.org/abs/2206.07314.

Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations, 2018. URL https://arxiv.org/abs/1803.07728.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples, 2015. URL https://arxiv.org/abs/1412.6572.

Sorin Grigorescu, Bogdan Trasnea, Tiberiu Cocias, and Gigel Macesanu. A survey of deep learning techniques for autonomous driving. *arXiv preprint arXiv:1910.07738*, 2019.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila 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, 2020. URL https://arxiv.org/abs/2006.07733.

Jie Gui, Tuo Chen, Jing Zhang, Qiong Cao, Zhenan Sun, Hao Luo, and Dacheng Tao. A survey on self-supervised learning: Algorithms, applications, and future trends. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9052–9071, 2024.

Rohit Gupta, Naveed Akhtar, Ajmal Mian, and Mubarak Shah. Contrastive self-supervised learning leads to higher adversarial susceptibility. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 14838–14846, 2023.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning, 2020. URL https://arxiv.org/abs/1911.05722.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16000–16009, June 2022.

Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty, 2019. URL https://arxiv.org/abs/1906.12340.

Qidong Huang, Xiaoyi Dong, Dongdong Chen, Yinpeng Chen, Lu Yuan, Gang Hua, Weiming Zhang, and Nenghai Yu. Improving adversarial robustness of masked autoencoders via test-time frequency-domain prompting, 2023. URL https://arxiv.org/abs/2308.10315.

Nils Hütten, Miguel Alves Gomes, Florian Hölken, Karlo Andricevic, Richard Meyes, and Tobias Meisen. Deep learning for automated visual inspection in manufacturing and maintenance: A survey of open-access papers. *Applied System Innovation*, 7(1):11, 2024. doi: 10.3390/asi7010011.

Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information, 2018. URL https://arxiv.org/abs/1804.08598.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pp. 448–456. pmlr, 2015.

Ziyu Jiang, Tianlong Chen, Ting Chen, and Zhangyang Wang. Robust pre-training by adversarial contrastive learning, 2020. URL https://arxiv.org/abs/2010.13337.

Minseon Kim, Jihoon Tack, and Sung Ju Hwang. Adversarial self-supervised contrastive learning. *Advances in neural information processing systems*, 33:2983–2994, 2020.

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

Raphaël Kulm. Is i-jepa an adversarially robust vision model? https://medium.com/@kulmluke/is-i-jepa-an-adversarially-robust-vision-model-110a697fdf0f, 2023. Accessed: 2025-10-14.

Peter Lorenz, Dominik Straßel, Margret Keuper, and Janis Keuper. Is robustbench/autoattack a suitable benchmark for adversarial robustness?, 2021. URL https://arxiv.org/abs/2112.01601.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. URL https://arxiv.org/abs/1711.05101.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks, 2019. URL https://arxiv.org/abs/1706.06083.

Sina Mohseni, Mandar Pitale, JBS Yadawa, and Zhangyang Wang. Self-supervised learning for generalizable out-of-distribution detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):5216–5223, Apr. 2020. doi: 10.1609/aaai.v34i04.5966. URL https://ojs.aaai.org/index.php/AAAI/article/view/5966.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks, 2016. URL https://arxiv.org/abs/1511.04599.

Giorgio Piras, Maura Pintor, Ambra Demontis, Battista Biggio, Giorgio Giacinto, and Fabio Roli. Adversarial pruning: A survey and benchmark of pruning methods for adversarial robustness. *Pattern Recognition*, 168:111788, 2025. doi: 10.1016/j.patcog.2025.111788. URL https://www.sciencedirect.com/science/article/pii/S0031320325004480.

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. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks, 2014. URL https://arxiv.org/abs/1312.6199.

Florian Tramèr, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On adaptive attacks to adversarial example defenses, 2020. URL https://arxiv.org/abs/2002.08347.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Andrew Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. Shows clean–robust accuracy trade-off.

Jonathan Uesato, Brendan O'Donoghue, Aaron van den Oord, and Pushmeet Kohli. Adversarial risk and the dangers of evaluating against weak attacks, 2018. URL https://arxiv.org/abs/1802.05666.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael I. Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning (ICML)*, 2019. TRADES: balances clean and robust accuracy.

SK Zhou, Hayit Greenspan, and Dinggang Shen. A review of deep learning in medical imaging: Imaging traits, technical challenges, and future directions. *arXiv preprint arXiv:2008.09104*, 2020.

A Appendix

A.1 Linear Probing Setup Details

We have two models based on the Visual Transformers ViT-H/14 architecture, and we consider two variations of each, as per the Linear Probing setup reflected in Assran et al. (2023).

- Supervised model ("google"): `google/vit-huge-patch14-224-in21k`, using the cross-entropy objective for classification
  - `bn off`: No batch normalization before the classification head
  - `bn on`: Batch normalization applied before the classification head
- I-JEPA model ("ijepa"): `facebook/ijepa_vith14_22k`, using the Image Joint-Embedding Predictive Architecture (I-JEPA) objective.
  - `last1`: Average pooling of the last layer patches before the classification head
  - `last4`: Concatenation of the average-pooled last four layer patches before the classification head

The training and evaluation for the linear probe setup are performed in Automatic Mixed Precision, so the evaluation results are slightly worse than those reflected in the Clean scores of the adversarial attack results, which are performed in Full Precision (FP32).

The batch normalization doesn't affect the classification results much. However, in I-JEPA models, the `last4` variant improves the performance on both CIFAR-100 and ImangeNet-100 by a few percentage points. Still, there is no reason to assume whether the `ijepa last4` variant provides better adversarial robustness as compared to the `ijepa last1` variant.

Based on these, we decide to use the Batch Norm variant for Google ViT (`google vit bn on`) and both the I-JEPA variants (`ijepa last1` and `ijepa last4`) for adversarial attacks.

Model	Variant	CIFAR-100 Top-1 Acc. %	ImageNet-100 Top-1 Acc. %
google vit	bn off	83.31	87.26
google vit	bn on	84.02	86.36
ijepa	last1	76.24	82.56
ijepa	last4	81.57	86.28

Table 2: Classification Accuracies of the different variations after training the linear probe.