Contrastive Self-Supervised Learning Leads to Higher Adversarial Susceptibility

Rohit Gupta[1], **Naveed Akhtar**[2], **Ajmal Mian**[2], **Mubarak Shah**[1]
[1]Center for Research in Computer Vision, University of Central Florida
[2] University of Western Australia
rohitg@knights.ucf.edu, {naveed.akhtar, ajmal.mian}@uwa.edu.au, shah@crcv.ucf.edu

Abstract

Contrastive self-supervised learning (CSL) has managed to match or surpass the performance of supervised learning in image and video classification. However, it is still largely unknown if the nature of the representations induced by the two learning paradigms is similar. We investigate this under the lens of adversarial robustness. Our analysis of the problem reveals that CSL has intrinsically higher sensitivity to perturbations over supervised learning. We identify the uniform distribution of data representation over a unit hypersphere in the CSL representation space as the key contributor to this phenomenon. We establish that this is a result of the presence of false negative pairs in the training process, which increases model sensitivity to input perturbations. Our finding is supported by extensive experiments for image and video classification using adversarial perturbations and other input corruptions. We devise a strategy to detect and remove false negative pairs that is simple, yet effective in improving model robustness with CSL training. We close up to 68% of the robustness gap between CSL and its supervised counterpart. Finally, we contribute to adversarial learning by incorporating our method in CSL. We demonstrate an average gain of about 5% over two different state-of-the-art methods in this domain.

Introduction

Contrastive Self-supervised Learning (CSL) (Chen et al. 2020a) is a widely adopted technique of self-supervised training of visual models (Radford et al. 2021), (Chen, Xie, and He 2021). It allows pre-training on unlabelled large-scale data prior to task-specific finetuning of the model. Since state-of-the-art CSL models in computer vision are performing equally well as the supervised models on popular benchmark datasets such as `ImageNet`, and make similar errors (Geirhos et al. 2020), it is a common perception that CSL models learn similar representations as their supervised counterparts.

In this work, we scrutinize the under-investigated question of similarity between supervised and CSL representations from the perspective of adversarial robustness. Surprisingly, our findings do not align well with the prevailing belief that both model types admit representations of similar nature. We find that CSL models are considerably inferior to supervised models in terms of adversarial robustness in image and video classification tasks. We show that this disparity also holds when equivalent augmentations and training schedules are applied to both supervised and self-supervised learning.

We investigate the reasons of higher adversarial vulnerability of CSL by examining the properties of the representation space induced by the contrastive loss. As illustrated in Fig. 1, Contrastive Learning (CL) brings similar pairs of data closer to each other in the learned representation space, and repels away the dissimilar pairs. In supervised CL, positive and negative pairs can be formed using known labels. For the self-supervised case, typically an instance in the dataset is treated as a negative for every other instance, and positives for an instance are generated using data augmentation.

Wang and Isola (2020) highlighted two key characteristics of CSL representation. (a) *Alignment*: positive pairs formed by data augmentation of a single instance are closer in the feature space, and (b) *uniformity*: a property that induces uniform distribution of instances in the representation space by repelling all other instances away from the anchor instance. The latter can be understood as an application of the principle of maximum entropy (Jaynes 1957) (colloquially referred to as *Occam's Razor*). Since class information about the instances is not available in self-supervised learning, it preserves maximum information about the data in the representation by inducing a uniform distribution of the training instances. Typically CSL representation lies on an n-dimensional hyper-sphere. As per the uniformity property, instances in the training dataset are roughly uniformly distributed over the surface of the hyper-sphere (Fig. 1b). However, this is sub-optimal from the perspective of adversarial robustness as this results in instances being close to the class boundaries and class clusters being spread out in the feature space. Contrastingly, in supervised learning, since the class labels are known, we can choose to only repel the instances from other classes to obtain tightly clustered classes in the feature space, achieving higher adversarial robustness.

A solution to this problem is to identify instance pairs in the unlabelled training data that belong to the same class and avoid using them as negative pairs for the contrastive loss. However, doing so is not trivial because data labels are unavailable in self–supervised learning. We propose an Adaptive False Negative Cancellation (Adaptive FNC) method that improves CSL training by gradually sifting out the

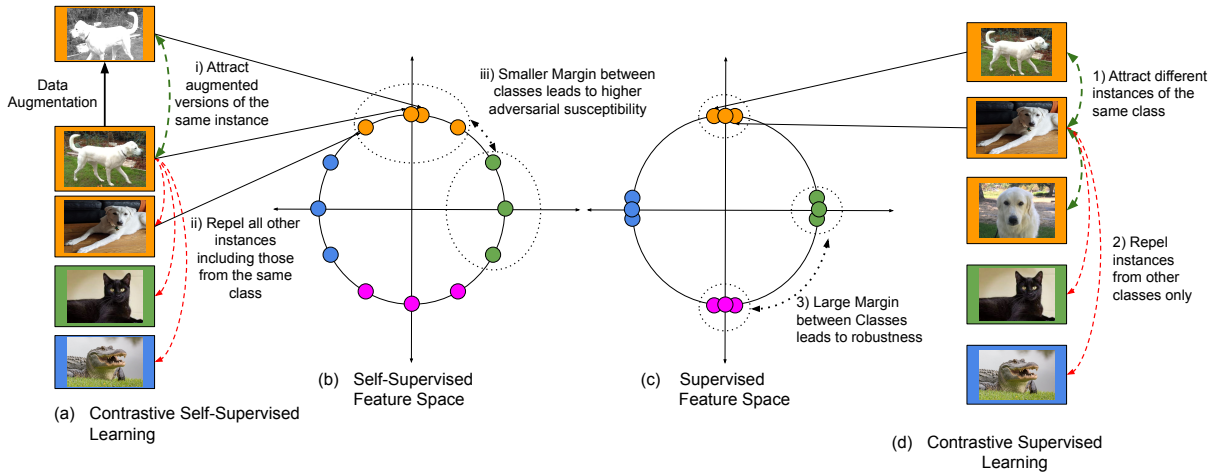


Figure 1: (a) In Contrastive Self-supervised Learning (CSL) the anchor instance forms a positive pair with its augmented version, while being uniformly repelled from all other instances, including instances of the same class, which results in (b) a representation space with large class clusters and small inter-class margins. (d) In Supervised Contrastive Learning (SupCon) all instances of a given class are attracted to each other, while instances of different classes are repelled, which allows for each class to occupy a compact region of the feature space (c), and as a result, has larger inter-class margins between class clusters. Lower inter-class margins in the self-supervised case lead to higher adversarial susceptibility. See Section for discussion.

likely false negative pairs. We demonstrate a significant adversarial robustness gain in the existing CSL methods RoCL and ACL with the proposed technique. Our key contributions are summarized below.

- We provide the first systematic evidence of higher sensitivity of CSL to input perturbations with extensive experimental verification for image and video classification.
- We establish a clear connection between adversarial susceptibility of CSL models and the uniformity of their representation. We identify false negative pairs in model training as the key reason of higher model sensitivity.
- Leveraging our insights, we devise a strategy to improve CSL robustness without adversarial training.
- We contribute to adversarial CSL by incorporating our findings into RoCL (Kim, Tack, and Hwang 2020) and ACL (Jiang et al. 2020a), achieving consistent performance gain against adversarial attacks.

Related Work

Contrastive learning: In self-supervised learning, Contrastive Learning (CL) is widely considered an effective paradigm. Within CL, SimCLR (Chen et al. 2020a) builds upon the prior work of MoCo (Momentum Contrastive learning) (He et al. 2020), Augmented Multiscale Deep InfoMax (AMDIM) (Bachman, Hjelm, and Buchwalter 2019), and Contrastive Predictive Coding (CPC) (Oord, Li, and Vinyals 2018) to develop its CL pipeline. The pipeline includes data augmentations and a projection head to align the learned network representation during training. While the performance of SimCLR has been lately matched or exceeded by MoCov2 (Chen et al. 2020d), MoCov3 (Chen, Xie, and He 2021) and SimCLRv2 (Chen et al. 2020b), the fundamental structure of the CL framework remains similar in these works. Contrastive learning has also been successfully extended to action classification in videos (Qian et al. 2021), (Dave et al. 2022), and image classification using Transformer architectures (Chen, Xie, and He 2021).

Many variants of CSL have been developed to improve its performance on clean data. MoCHi (Kalantidis et al. 2020) generates synthetic hard negatives in representation space, and NNCLR (Dwibedi et al. 2021) dynamically mines positive examples from different instances. Debiased CSL (Chuang et al. 2020) modifies the loss function to account for the presence of negative pairs formed by different instances belonging to the same class. FNC (Huynh et al. 2022) and IFND (Chen et al. 2022) attempt to directly identify and remove negative pairs during training. This concept is related to our approach, however, we focus on improving adversarial robustness whereas these prior works focus on improving accuracy on clean data.

Even though SwAV (Caron et al. 2020) does not use contrastive loss, it preserves the 'alignment' property of its representation by clustering the augmented versions of instances. Moreover, it is also able to preserve the 'uniformity' property by enforcing an explicit equi-partitioning constraint over its representation space. Recent non-CSL methods include SimSiam (Chen and He 2021), BYOL (Grill et al. 2020) and DINO (Caron et al. 2021), which are self-distillation based methods. These techniques do not use negative contrastive pairs and hence are not directly affected by the weaknesses of CSL.

Owing to the promising performance of CL, recent works have also focused on exploring the unique properties of contrastive learning. Geirhos et al. (Geirhos et al. 2020) found that such models produce results similar to those learned

with supervision. Xiao et al. (2021) showed that the choice of the best data augmentation method for self-supervised training depends on the specific dataset. Purushwalkam and Gupta (2020) claimed that CL results in superior occlusion-invariant representations, while Wang and Isola (2020) analyzed CL by studying the alignment and uniformity properties of feature distribution. These properties are claimed to endow more discriminative power to the models. The uniformity property of CL is also discussed by Chen, Luo, and Li (2021), who refer to it as the 'distribution' property. Wang and Liu (2021) built a relationship between the uniformity and the temperature hyper-parameter of the loss function.

Robustness and self-supervision: In prior art on robustification of supervised learning, self-supervision is considered as a helpful tool. Hendrycks et al. (2019) found that adversarial robustness of supervised models can be improved by adding an additional self-supervised task in a multi-task approach. Similarly, Carmon et al. (2019) also found that using additional unlabeled data improves adversarial robustness of the model. Chen et al. (2020c) also developed robust versions of pretext-based self-supervised learning tasks and demonstrated that this, along with robust fine-tuning of the model, results in significant improvement in the robustness relative to the baseline adversarial training.

Adversarial training for self-supervised models: Efforts have also been made for adversarial training of self-supervised learning. Kim, Tack, and Hwang (2020) developed an instance-based adversarial attack for contrastive self-supervised training, and used it for adversarial training. The concurrent work by Jiang et al. (2020b) develops an adversarial CL framework that is claimed to surpass prior self-supervised learning methods in robustness as well as accuracy on clean data. Ho and Nvasconcelos (2020) created a generalized formulation of AdvProp training (Xie et al. 2020) applicable to self-supervised learning, with the goal to increase accuracy on clean data. Adversarial training increases robustness of CL models, but incurs a significant training cost. Hence, we contribute to CSL robustness without resorting to adversarial training. We also show that adversarial training can further benefit from our technique.

Adversarial Susceptibility of CSL

The popular contrastive self-supervised representation learning strategy, e.g., used by SimCLR (Chen et al. 2020a), learns a representation space from unlabeled data. It samples the so-called 'positive pairs' by applying independent random transformations to an original sample (a.k.a. anchor). The positive pairs are expected to have representations similar to the anchor. The 'negative pairs' are formed by pairing the anchor with other original instances. If x represents the anchor, x^+ represents the positive pairing generated by augmentation and each x_i^- represents another instance in the training batch, then the contrastive loss, $\mathcal{L}_c(x)$, is given by:

$$\mathcal{L}_c(x) = \mathop{\mathbb{E}}_{x,x^+,\{x_i^-\}_i^N}\Big[-\log\frac{sim(x,x^+)}{sim(x,x^+)+\sum_i^N sim(x,x_i^-)}\Big] \quad (1)$$

For brevity in Eq.(1) we define $sim(x,y)$ as $e^{f(x)^T\cdot f(y)/\tau}$, which is the normalized temperature scaled similarity between the representations of inputs x and y. Here, $f(x)$ is the representation of a given input, which is constrained by design to lie on a d dimensional hypersphere.

In order to link CSL to adversarial susceptibility, we make following three claims:

1. Presence of false negative pairs in CSL leads to *instance-level* uniformity
2. Instance level uniformity implies weaker separation of classes in the feature space (lower ratio of average inter-class to intra-class distance)
3. Weaker separation of classes results in higher adversarial susceptibility

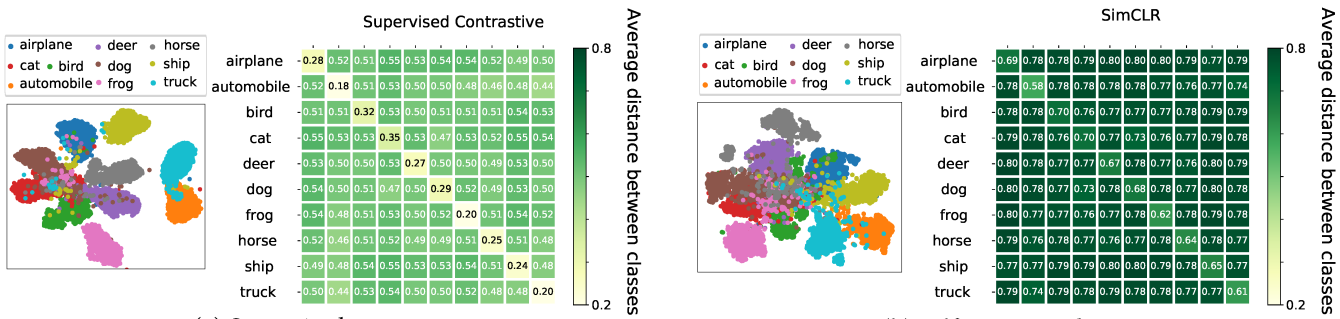
Firstly, since all instances in CSL loss repel each other, it is likely (illustrated in Figure 1) that instance level uniformity is the stable equilibrium to minimize the loss. However, prior work (Wang and Isola 2020) also provides an analytical proof that in the limit of large batch size ($N \to \infty$) the contrastive loss is equivalent to reducing:

$$\mathcal{L}'_c(x) \equiv -\frac{1}{\tau}\mathop{\mathbb{E}}_{x,x^+} sim(x,x^+) + \mathop{\mathbb{E}}_{x,x^-}\log sim(x^-,x). \quad (2)$$

Here, the first term ensures *alignment* of positive pairs, whereas the second term ensures *uniformity* of instances over the hypersphere. Note that in the case of supervised learning, since instances of the same class are not repelled, there is no instance level uniformity. Rather, the uniformity term applies at the *class-level*.

Which brings us to our second claim. We know from geometry that the surface area of a d-dimensional sphere is given by: $\mathrm{A}(d) = \Big(2\pi^{\frac{d-1}{2}}/\Gamma(\frac{d-1}{2})\Big)$, where $\Gamma(x)$ is the gamma function. This surface area is divided very differently in supervised and self-supervised learning. Since the classes as uniformly spread out, the distance between a class center and the nearest boundary should be proportional to $A_c = \frac{\mathrm{A}(d)}{C}$, where C is the number of classes in the dataset. In the supervised case, as training progresses, instances of a class will come closer and closer to each other. We choose to indicate this minimal intra-class separation by s_{min}. Whereas in the self-supervised case, the intra-class margin has to be higher than $\frac{\mathrm{A}(d)}{N}$ due to instance level uniformity. If we define a measure of class separation in terms of the ratio of the average inter-class distance to the average intra-class distance, we would expect $\rho_{SupCon} = \frac{A_c}{s_{min}}$ and $\rho_{CSL} = \frac{A_c \times N}{C}$. Since $s_{min} \to 0$ and $N/C \in$ finite rationals $\mathbb{Q}$, which implies $\rho_{SupCon} > \rho_{CSL}$. In Section we also provide empirical evidence to show that intra-class distances are much smaller for the supervised models than for CSL.

Finally, we come to our third claim, which asserts that classifiers with stronger separation of classes in feature space are more robust. This follows intuitively from the definition of adversarial examples. Adversarial examples exist because clean examples from the dataset lie very close



(a) Supervised (b) Self-Supervised

Figure 2: tSNE Visualization of representation space and average inter- and intra- class distances for CIFAR-10 instance pairs obtained with (a) Supervised, and (b) Self-Supervised contrastive learning. Average ratio of inter-class distances relative to intra-class distances is much lower for the Self-Supervised model (1.19×) than for Supervised (1.98×), which leads to higher adversarial susceptibility for the self-supervised model.

Pre-Training	**FGSM-ℓ_∞**			**PGD-ℓ_2**		**PGD-ℓ_1**	**AutoAttack-ℓ_∞**
	ϵ = **0**	ϵ = **8/255**	ϵ = **16/255**	ϵ = **0.1**	ϵ = **0.25**	ϵ = **8.0**	ϵ = **1/255**
CIFAR-10							
Supervised (Cross entropy)	95.4	29.6 ↓69%	20.1 ↓79%	31.2 ↓67%	15.7 ↓84%	16.5 ↓83%	18.6 ↓81%
Supervised (Contrastive)	95.5	38.8 ↓59%	31.8 ↓67%	34.2 ↓64%	18.4 ↓81%	20.7 ↓78%	24.3 ↓75%
Self-Supervised (Contrastive)	92.7	26.8 ↓71%	13.4 ↓86%	20.9 ↓77%	8.3 ↓91%	11.5 ↓88%	13.7 ↓85%
CIFAR-100							
Supervised (Cross entropy)	74.9	14.3 ↓81%	8.4 ↓89%	23.1 ↓69%	11.5 ↓85%	12.1 ↓84%	13.6 ↓82%
Supervised (Contrastive)	76.3	12.6 ↓83%	6.7 ↓91%	21.9 ↓71%	9.2 ↓88%	13.4 ↓82%	11.9 ↓84%
Self-Supervised (Contrastive)	68.9	9.40 ↓87%	3.0 ↓96%	13.7 ↓80%	4.4 ↓94%	6.8 ↓90%	8.7 ↓88%

Table 1: Supervised and CSL CIFAR models are trained with similar training setups and their robustness is compared for FGSM, `AutoAttack` (both ℓ_∞) and PGD variants (ℓ_1 & ℓ_2).

to decision boundaries in feature space and can be pushed across them with a small perturbation. Prior works such as DeepFool (Moosavi-Dezfooli, Fawzi, and Frossard 2016) define classifier robustness in terms of the average minimum perturbation magnitude required to shift the labels for the dataset. Since deep classifiers are typically smooth, a larger magnitude of perturbation required in pixel space indicates higher distance of instance from nearest decision boundary.

Empirical Verification of Susceptibility

In order to empirically verify our finding of higher susceptibility for CSL, we compare robustness of supervised and self-supervised image and video classification models. First, we train models with identical architectures and comparable training policies on the CIFAR-10 and CIFAR-100 datasets in order to isolate the effect of CSL training on robustness. Second, we demonstrate that our finding of adversarial susceptibility of CSL also extends to ImageNet pretrained models. Third, we also demonstrate that to analyze the robustness of models to image transformations e.g blurring and noise addition. Finally, we move to the task of video classification and demonstrate that our observations also apply to models trained on UCF101 and HMDB51 datasets, two widely used benchmarks for action recognition.

For comparing the robustness of models (which can have varying clean accuracy) we use the drop in accuracy relative to the clean accuracy:

$$\mathcal{P}_{Drop} = \frac{\mathcal{P}(y|x) - \mathcal{P}(y|x+\Delta x)}{\mathcal{P}(y|x)}, \tag{3}$$

where $\mathcal{P}(y|x)$ is the model accuracy for clean data and $\mathcal{P}(y|x+\Delta x)$ is the accuracy with perturbed data. x and y represent data and correct label respectively.

Susceptibility of CIFAR Models

We perform controlled experiments with supervised cross-entropy, supervised contrastive learning (Khosla et al. 2020) and CSL using CIFAR-10 and CIFAR-100 datasets. While our argument on higher sensitivity of contrastive self-supervised learning is best verified using a weaker attack like FGSM, we also investigate the adversarial susceptibility of models to stronger attacks, e.g., Projected Gradient Descent (PGD) (Madry et al. 2018) and AutoAttack (Croce and Hein 2020b). AutoAttack is a composite of 4 attacks: untargeted and targeted step-size free Automatic PGD, Fast Adaptive Boundary Attack (Croce and Hein 2020a) and the

Model	Method		FGSM		PGD-ℓ_∞		PGD-ℓ_2	AutoAttack-ℓ_∞
		ϵ = 0	ϵ = .25/255	ϵ = 1/255	ϵ = .25/255	ϵ = 1/255	ϵ = 0.5	ϵ = .25/255
ResNet50	Supervised	76.71	38.20 ↓50%	11.53↓85%	28.22 ↓63%	0.65 ↓99%	11.3 ↓85%	16.16↓79%
	SimCLR	68.95	24.33 ↓65%	8.85 ↓87%	10.89 ↓84%	0.24 ↓100%	5.2 ↓92%	4.41↓94%
	SwAV	75.34	23.35 ↓69%	5.95 ↓92%	11.73 ↓84%	0.20 ↓100%	4.1 ↓95%	10.81↓86%
	BYOL	74.56	39.40 ↓47%	13.50↓82%	32.61 ↓56%	1.89 ↓97%	12.51↓83%	17.34↓77%
ViT-B/16	Supervised	81.07	61.44 ↓24%	43.38↓46%	59.57 ↓27%	27.53↓66%	42.2 ↓48%	46.99 ↓42%
	MoCoV3	76.66	46.69 ↓39%	13.21↓83%	39.89 ↓48%	2.42 ↓97%	19.6 ↓75%	27.88 ↓64%
	DINO	77.99	58.36 ↓25%	22.17↓72%	57.95 ↓26%	14.14↓82%	36.6 ↓53%	51.46 ↓34 %

Table 2: ImageNet top-1 accuracy for various pre-trained models under FGSM, PGD and AutoAttack attacks. Percentage drop relative to clean input accuracy is indicated as ↓x%. Contrastive Self-Supervised models show higher drops.

black-box Square Attack (Andriushchenko et al. 2020). In order to ensure a fair comparison, all our ResNet50 models are trained for 1,000 epochs each, which ensures full convergence. We also use the same data augmentation strategy for all models, with the minor difference of using weaker color jittering for the supervised cross-entropy loss, as suggested by (Chen et al. 2020a). To ensure reliability of results, they are averaged over 5 training runs. The standard deviation for all settings and models was less than $< 0.5\%$,

The results of our experiments are summarized in Table 1, which suggest that supervised models with contrastive loss are more resilient to adversarial manipulation as compared to their self-supervised counterparts. In Fig. 2, we provide tSNE visualisation of the representations for supervised and self-supervised CIFAR-10 models learned under the contrastive loss in Table 1. As can be observed from the respective class heatmaps, the ratio of inter-class to intra-class margin is much higher for supervised ($1.98\times$) than in the self-supervised case ($1.19\times$). Clearly, the supervised model is able to separate the classes better than the self-supervised model. This observation is in line with our finding that uniform representations in self-supervised contrastive learning renders the model more sensitive to input perturbations.

Susceptibility of ImageNet Models

In order to verify our results for widely deployed pre-trained ImageNet models, we focus on two popular architectures: ResNet50 and Vision Transformer (ViT-B). We compare the robustness of supervised models against models trained with self-supervision techniques. We select both CSL and non-CSL self-supervision methods in order to demonstrate the validity of our argument. SimCLR (Chen et al. 2020a), a simple CSL method, and MoCov3, a state of the art technique which utilizes an additional momentum encoder and achieves better results are our primary focus. We also include SwAV (Caron et al. 2020), which does not use contrastive loss, however, it also preserves the uniformity property of representations, which, according to our analysis in Section , is the primary cause of higher adversarial susceptibility. Finally, we test BYOL and DINO, which are a self-supervised method based on self-distillation and do not utilize negative pairs. We maintain architectural similarity between different techniques to ensure comparable results. Along with adversarial attacks, we also study the robustness of models w.r.t. other transformations e.g., adding noise, blurring, simulated fog etc., using ImageNet-C dataset.

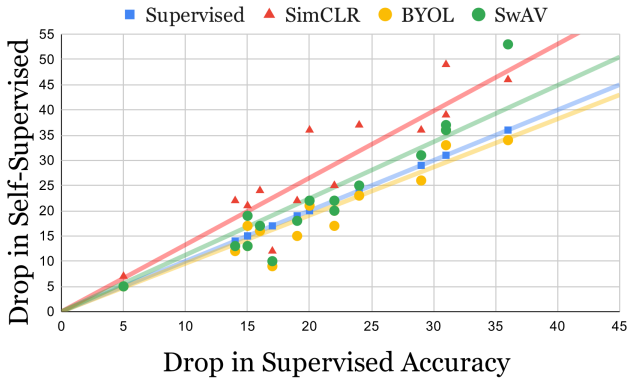


Figure 3: Relative accuracy drop (ResNet50) due to ImageNet-C corruptions. CSL (SimCLR) has bigger drops than Supervised or SSL without negative pairs (BYOL).

Susceptibility to Adversarial Perturbations: We test the adversarial robustness of popular pre-trained ImageNet Classifiers against 3 different attacks: FGSM, PGD and AutoAttack. The results for these experiments are provided in Table 2. In the table, we use FGSM by varying its perturbation scale ϵ from the set $0, 0.25, 1$, where 0 indicates clean images. The image dynamic range is $[0, 255]$. For the reported top-1 accuracy, percentage reductions for CSL methods SimCLR, MoCov3 and SwAV are much larger than for the supervised model. The difference is particularly large for the weaker perturbations, which indicates the higher sensitivity of the model predictions. The results align well with the insights in Section . The observation also holds for the variants of the stronger PGD attack. We provide results for the standard ℓ_∞ and ℓ_2 variants of the algorithm, performing 40 iterations, which is commonly used in the literature. Table 2 points to the higher relative adversarial sensitivity of the self-supervision models. DINO and BYOL which do not utilize negative pairs during training, demonstrates higher robustness than the CSL methods SimCLR and MoCov3, which provides further evidence for our hypothesis. We also note that for the pre-trained checkpoints, ViT models are more robust than ResNet. This observation has previously been reported in the literature (Naseer et al. 2021). However as other works (Pinto, Torr, and K Dokania 2022; Bai et al. 2021) have suggested that the root cause cannot be attributed to differences in architecture, we do not make any claim

Pre-Training	$\epsilon = 0$	FGSM-ℓ_∞ $\epsilon = 1/255$	FGSM-ℓ_∞ $\epsilon = 2/255$	FGSM-ℓ_∞ $\epsilon = 4/255$	PGD-ℓ_∞ $\epsilon = 1/255$	PGD-ℓ_∞ $\epsilon = 2/255$	AutoAttack-ℓ_∞ $\epsilon = 1/255$
UCF101							
Supervised	59.4	26.6 ↓55%	12.2 ↓79%	3.9 ↓93%	15.6 ↓74%	5.2 ↓91%	12.6 ↓79%
TCLR (Dave et al. 2022)	75.5	10.8 ↓86%	6.1 ↓92%	3.3 ↓96%	6.30 ↓92%	3.1 ↓96%	4.00 ↓95%
CVRL (Qian et al. 2021)	60.2	6.00 ↓90%	3.0 ↓95%	1.6 ↓97%	4.70 ↓92%	1.9 ↓97%	3.70 ↓94%
HMDB51							
Supervised	24.8	12.2 ↓51%	6.4 ↓74%	3.0 ↓88%	7.2 ↓71%	2.7 ↓89%	6.2 ↓75%
TCLR (Dave et al. 2022)	47.6	5.7 ↓77%	2.7 ↓89%	1.7 ↓93%	2.7 ↓89%	2.0 ↓92%	2.0 ↓92%
CVRL (Qian et al. 2021)	35.2	3.7 ↓85%	1.7 ↓93%	0.5 ↓98%	2.2 ↓91%	1.0 ↓96%	1.7 ↓93%

Table 3: Top-1 accuracy for pre-trained UCF101 and HMDB51 video classification under FGSM, PGD and AutoAttack attacks. Percentage drop relative to clean data accuracy is indicated as ↓x%. CSL methods TCLR and CVRL have lower robust accuracy than the supervised model despite heavily outperforming supervised learning in clean accuracy.

about the relative robustness of transformers and CNNs.

Susceptibility to Natural Image Corruptions: We also employ the ImageNet-C dataset (Hendrycks and Dietterich 2019) to analyze the robustness of models to more primitive transformations, e.g., blurring and noise addition. ImageNet-C includes these perturbations at 5 increasing distortion levels. However, the lowest level is the most relevant to our analysis because we are concerned with the higher local sensitivity of models. We summarize those results in Figure. 3. On average SimCLR & SwAV experience a 33% and 12% higher drop in accuracy relative to the supervised model, while BYOL performs about equally as well as supervised. The results establish higher overall sensitivity of the contrastive self-supervised models for 15 image corruption types. The few cases in which SimCLR shows low sensitivity to corruptions are Brightness and Contrast jittering, which are used as augmentations during its training. SwAV is relatively more robust than SimCLR to non-adversarial transformations, which is a natural consequence of its ability to 'cluster' positive samples for a class.

Susceptibility of Video Classification

To establish that our observations also hold for different types of models, we perform analysis for action recognition, which is a video classification task. Recently, action recognition techniques have started to exploit contrastive learning. This opens up the avenue of adversarial robustness analysis for the problem. We use FGSM, PGD and AutoAttack based attacks here and the results are consistent across attacks. We employ an 18-layer R-(2+1)-D model in our experiments. One variant is trained with supervised cross-entropy loss, and other two are trained using contrastive self-supervised learning methods, TCLR (Dave et al. 2022) and CVRL (Qian et al. 2021). We summarize our results on two datasets (UCF101 and HMDB51) in Table 3 which shows that the CSL models also gets affected more strongly by the attack as compared to the supervised models. This is true despite self-supervised models considerably outperforming the supervised model on clean inputs.

Enhancing CSL Robustness

We have thoroughly established that self-supervised contrastive learning is more sensitive to adversarial inputs than supervised learning. Our analysis in Section points to the presence of false negative instances in the training data as the major cause of this higher sensitivity. Thus, detecting and removing those can potentially improve the model robustness. In this section we seek to demonstrate that the CSL process itself can be modified to address the issue. Our modifications can further also be applied to adversarial training methods for CSL and achieve gains in robustness without adding any significant computational overhead to the process.

Adaptive False Negative Cancellation

Identifying instances that belong to the same semantic class is not straightforward in the absence of label information. By definition, false negatives must share similar features as the true positives. This suggests that we can decide on a suspect false negative by measuring the similarity between a sample's representation to that of the anchor in our mini-batch. Recall from Section that two prior works have proposed variants of CSL utilizing false negative removal strategies with the goal of improving performance on clean data. False Negative Cancellation (FNC) (Huynh et al. 2022) proposes cancelling all negatives which have a cosine similarity higher than a fixed threshold w.r.t. to the anchor and also additionally top-k similar negatives from each batch. Incremental False Negative Detection (IFND) (Chen et al. 2022) on the other hand utilizes offline unsupervised clustering to propagate psuedo-labels. The key benefit of FNC is that it does not increase the training time over simple CSL significantly, however unlike IFND, it utilizes a fixed threshold and k value and as a result does not adapt to the shifting features learned by the model as training progresses.

In order to maintain the low computational overhead of FNC, while also accounting for the improvement in representation quality as training progresses, we propose Adaptive False Negative Cancellation (Adaptive FNC). In Adaptive FNC, we replace the standard CSL objective with a modified objective $\mathcal{L}_a(x)$:

$$\mathcal{L}_a(x) = \mathbb{E}\Big[-\log \frac{sim(x, x_i^+)}{sim(x, x_i^+) + \sum_i^{N^-} sim(x, x_i^-)} \Big] \quad (4)$$

Here, the set of selected positive instances $\{x_i^+\}$ and filtered negative instances $\{x_i^-\}$ is chosen on the basis of similarity in the feature space. A dynamic similarity threshold

Pre-Training	Attacks →	FGSM-ℓ_∞		AutoAttack-ℓ_∞	
	ϵ = **0**	ϵ = **8/255**	ϵ = **16/255**	ϵ = **1/255**	ϵ = **2/255**
Supervised Contrastive (SupCon)	95.5	38.8 ↓59%	31.8 ↓67%	24.3 ↓74%	11.5 ↓88%
Self-Supervised Contrastive (SimCLR)	92.7	26.8 ↓71%	13.4 ↓86%	13.7 ↓85%	4.3 ↓95%
SimCLR + FNC (Huynh et al. 2022)†	94.9	28.6 ↓70% ↑**8%**	16.3 ↓83% ↑**16%**	14.7 ↓85% ↑**1%**	6.6 ↓93% ↑**27%**
SimCLR + IFND (Chen et al. 2022)†	95.1	29.2 ↓69% ↑**16%**	18.8 ↓80% ↑**32%**	15.1 ↓84% ↑**9%**	7.1 ↓93% ↑**29%**
Ours (SimCLR + Adaptive FNC)					
Fixed Threshold ($\rho = 0.9$)	93.1	30.1 ↓68% ↑**25%**	21.3 ↓77% ↑**53%**	17.4 ↓81% ↑**25%**	7.5 ↓92% ↑**43%**
Adaptive Threshold ($\rho = 0.9 \rightarrow 0.5$)	93.2	31.3 ↓66% ↑**42%**	21.6 ↓76% ↑**53%**	19.1 ↓79% ↑**55%**	8.6 ↓90% ↑**70%**
Adaptive Threshold ($\rho = 0.99 \rightarrow 0.7$)	**93.5**	**33.6** ↓64% ↑**58%**	**24.9** ↓73% ↑**68%**	**19.3** ↓79% ↑**58%**	**8.7** ↓90% ↑**72%**

Table 4: Robustness improvement with our method. Top-1 accuracy for CIFAR-10 under FGSM and AutoAttack. The first two rows provide results without false negative removal. Drop in accuracy under attack is reported as ↓x%, percentage of gap closed w.r.t. supervised contrastive learning is indicated as ↑x%. Best Result closes >58% of the gap. †- re-implementation.

Pre-Training		PGD-ℓ_∞		PGD-ℓ_2	PGD-ℓ_1	AutoAttack-ℓ_∞	
	$\epsilon = 0$	$\epsilon = \frac{8}{255}$	$\epsilon = \frac{16}{255}$	$\epsilon = 0.25$	$\epsilon = 12$	$\epsilon = \frac{8}{255}$	$\epsilon = \frac{16}{255}$
Supervised	95.5	0.0	0.0	24.8	25.4	0.0	0.0
Self-Supervised	92.7	0.0	0.0	17.1	21.1	0.0	0.0
RoCL (Kim, Tack, and Hwang 2020)	86.0	43.6	11.4	70.9	80.0	40.8	11.2
Ours *(Augmented-RoCL)*	87.9	45.9 ↑**5.3%**	13.2 ↑**15.8%**	72.8 ↑**2.7%**	82.1 ↑**2.6%**	43.1 ↑**5.6%**	12.1 ↑**8.0%**
ACL (Jiang et al. 2020b)	86.2	41.2	12.1	72.3	80.7	39.8	10.2
Ours *(Augmented-ACL)*	87.9	42.5 ↑**3.1%**	13.2 ↑**9.5%**	75.9 ↑**5.0%**	83.5 ↑**3.5%**	41.3 ↑**3.7%**	10.8 ↑**5.5%**

Table 5: Top-1 accuracy of adversarially trained CIFAR-10 models under PGD attack and AutoAttack. Attack strength ϵ is expressed in terms of ℓ_∞, ℓ_2 and ℓ_1 norms. The first two rows provide results without adversarial training. Robust models are trained with PGD ℓ_∞ adversary. Percentage performance gain of our false negative removal augmented methods over adversarially trained RoCL (median gain across attacks is 5.5) and ACL (median gain 4.4%) is indicated as ↑x% .

ρ is used, and instances with similarity higher than ρ are selected into $\{x_i^+\}$ and the rest form $\{x_i^-\}$. Since the quality of features is initially low, we begin with a high value of ρ to avoid spurious detections and slowly decrease it as training progresses and similarity scores become more reliable. Removing false negative pairs from the training objective should lead to tighter class clusters and more adversarially robust models. Adaptive FNC is also efficient as detection using the threshold is $O(N)$ during loss computation and does not increase the computational complexity of the training process. Note that unlike prior work FNC we do not use a top-K heuristic to sample additional false negatives. We also avoid False Negative Attraction as detection of false negative pairs during initial phase of training is not accurate and using them as positive pairs negatively impacts robustness.

Experiments: We carry out an ablation study with $\rho_{initial} \in \{0.99, 0.9\}$ and $\rho_{final} \in \{0.5, 0.7\}$ in order to determine the best possible setting for our Adaptive FNC method. The results for the two extreme cases are presented in Table 4. A simple baseline with fixed threshold ($\rho_{initial} = \rho_{final} = 0.9$) is also presented to illustrate the benefits of adaptive FNC. Prior works FNC (Huynh et al. 2022) and IFND (Chen et al. 2022), focus on improving performance on clean data outperform our method in clean accuracy. However, Adaptive FNC provides substantial improvement in robust accuracy under FGSM and AutoAttack. The improvement here is without adversarial training and with minimal overhead.

Augmenting Adversarial CSL

Adversarial learning is a widely adopted paradigm for learning robust models in the supervised learning domain (Akhtar et al. 2021). We demonstrate that our Adaptive FNC technique can readily augment adversarial learning in the CSL domain. To that end, we enhance Robust Contrastive Learning (RoCL) (Kim, Tack, and Hwang 2020) and Adversarial Contrastive Learning (ACL) (Jiang et al. 2020a) methods with our Adaptive FNC technique. Here, it is also pertinent to mention that referring to (Kim, Tack, and Hwang 2020), (Carmon et al. 2019), (Chen et al. 2020c), Hendrycks et al. (Hendrycks et al. 2019) alluded to the idea that self-supervision can help in adversarial robustness. Our previous findings provide evidence against this idea for CSL methods. This makes our contribution towards the enhancement of adversarial contrastive learning even more relevant. Discussion on the details of enhancing the RoCL and ACL methods with our technique are provided in the supplementary material.

Experiments: To evaluate the performance gain achieved by A-RoCL and A-ACL, we followed (Kim, Tack, and Hwang 2020) and performed adversarial training with PGD ℓ_∞-norm bounded adversary. The model robustness is evaluated for ℓ_∞, ℓ_1, ℓ_2 PGD and ℓ_∞ AutoAttack. Results are averaged over 5 training runs and summarized in Table 5. Our Augmented-RoCL and Augmented-ACL consistently improve performance gain over RoCL (Kim, Tack, and Hwang 2020) and ACL (Jiang et al. 2020a) baselines.

Conclusion

In this paper, we investigate the adversarial susceptibility of Contrastive Self Supervised Learning (CSL) trained models. We empirically verify that the use of negative pairs during CSL training is linked to susceptibility. This link is then used to enhance the robustness of CSL and adversarial CSL.

Acknowledgements

This work was supported in part by the Defense Advanced Research Projects Agency (DARPA) under Agreement HR00112090095. Dr. Naveed Akhtar is the recipient of Office of National Intelligence, National Intelligence Postdoctoral Grant (project number NIPG-2021-001) funded by the Australian Government. Professor Ajmal Mian is the recipient of an Australian Research Council Future Fellowship Award (project number FT210100268) funded by the Australian Government.

The authors would also like to thank Ishan Dave for fruitful discussions on video contrastive learning methods.

References

Akhtar, N.; Mian, A.; Kardan, N.; and Shah, M. 2021. Advances in adversarial attacks and defenses in computer vision: A survey. *IEEE Access*, 9: 155161–155196.

Andriushchenko, M.; Croce, F.; Flammarion, N.; and Hein, M. 2020. Square attack: a query-efficient black-box adversarial attack via random search. In *European Conference on Computer Vision*, 484–501. Springer.

Bachman, P.; Hjelm, R. D.; and Buchwalter, W. 2019. Learning Representations by Maximizing Mutual Information Across Views. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d Alché-Buc, F.; Fox, E.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 32.

Bai, Y.; Mei, J.; Yuille, A. L.; and Xie, C. 2021. Are Transformers more robust than CNNs? *Advances in Neural Information Processing Systems*, 34: 26831–26843.

Carmon, Y.; Raghunathan, A.; Schmidt, L.; Duchi, J. C.; and Liang, P. S. 2019. Unlabeled data improves adversarial robustness. *Advances in Neural Information Processing Systems*, 32.

Caron, M.; Misra, I.; Mairal, J.; Goyal, P.; Bojanowski, P.; and Joulin, A. 2020. Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. In *Advances in Neural Information Processing Systems*, volume 33.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9650–9660.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020a. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PMLR.

Chen, T.; Kornblith, S.; Swersky, K.; Norouzi, M.; and Hinton, G. E. 2020b. Big Self-Supervised Models are Strong Semi-Supervised Learners. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M. F.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 22243–22255.

Chen, T.; Liu, S.; Chang, S.; Cheng, Y.; Amini, L.; and Wang, Z. 2020c. Adversarial robustness: From self-supervised pre-training to fine-tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 699–708.

Chen, T.; Luo, C.; and Li, L. 2021. Intriguing Properties of Contrastive Losses. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*, volume 34, 11834–11845.

Chen, T.-S.; Hung, W.-C.; Tseng, H.-Y.; Chien, S.-Y.; and Yang, M.-H. 2022. Incremental False Negative Detection for Contrastive Learning. In *International Conference on Learning Representations*.

Chen, X.; Fan, H.; Girshick, R.; and He, K. 2020d. Improved Baselines with Momentum Contrastive Learning. arXiv:2003.04297.

Chen, X.; and He, K. 2021. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15750–15758.

Chen, X.; Xie, S.; and He, K. 2021. An Empirical Study of Training Self-Supervised Vision Transformers. arXiv:2104.02057.

Chuang, C.-Y.; Robinson, J.; Lin, Y.-C.; Torralba, A.; and Jegelka, S. 2020. Debiased contrastive learning. *Advances in neural information processing systems*, 33: 8765–8775.

Croce, F.; and Hein, M. 2020a. Minimally distorted adversarial examples with a fast adaptive boundary attack. In *International Conference on Machine Learning*, 2196–2205. PMLR.

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

Dave, I.; Gupta, R.; Rizve, M. N.; and Shah, M. 2022. Tclr: Temporal contrastive learning for video representation. *Computer Vision and Image Understanding*, 219: 103406.

Dwibedi, D.; Aytar, Y.; Tompson, J.; Sermanet, P.; and Zisserman, A. 2021. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9588–9597.

Geirhos, R.; Narayanappa, K.; Mitzkus, B.; Bethge, M.; Wichmann, F. A.; and Brendel, W. 2020. On the surprising similarities between supervised and self-supervised models. *NeurIPS workshop on Shared Visual Representations in Human and Machine Intelligence*.

Grill, J.-B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P.; Buchatskaya, E.; Doersch, C.; Avila Pires, B.; Guo, Z.; Gheshlaghi Azar, M.; Piot, B.; Kavukcuoglu, K.; Munos, R.; and Valko, M. 2020. Bootstrap Your Own Latent - A New Approach to Self-Supervised Learning. In Larochelle,

H.; Ranzato, M.; Hadsell, R.; Balcan, M. F.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 21271–21284.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9729–9738.

Hendrycks, D.; and Dietterich, T. 2019. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. In *International Conference on Learning Representations*.

Hendrycks, D.; Mazeika, M.; Kadavath, S.; and Song, D. 2019. Using Self-Supervised Learning Can Improve Model Robustness and Uncertainty. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d Alché-Buc, F.; Fox, E.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 32.

Ho, C.-H.; and Nvasconcelos, N. 2020. Contrastive Learning with Adversarial Examples. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M. F.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 17081–17093.

Huynh, T.; Kornblith, S.; Walter, M. R.; Maire, M.; and Khademi, M. 2022. Boosting contrastive self-supervised learning with false negative cancellation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2785–2795.

Jaynes, E. T. 1957. Information theory and statistical mechanics. *Physical review*, 106(4): 620.

Jiang, Z.; Chen, T.; Chen, T.; and Wang, Z. 2020a. Robust pre-training by adversarial contrastive learning. *Advances in Neural Information Processing Systems*, 33: 16199–16210.

Jiang, Z.; Chen, T.; Chen, T.; and Wang, Z. 2020b. Robust Pre-Training by Adversarial Contrastive Learning. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M. F.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 16199–16210.

Kalantidis, Y.; Sariyildiz, M. B.; Pion, N.; Weinzaepfel, P.; and Larlus, D. 2020. Hard negative mixing for contrastive learning. *Advances in Neural Information Processing Systems*, 33: 21798–21809.

Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; and Krishnan, D. 2020. Supervised Contrastive Learning. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M. F.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 18661–18673.

Kim, M.; Tack, J.; and Hwang, S. J. 2020. Adversarial Self-Supervised Contrastive Learning. In *Thirty-fourth Conference on Neural Information Processing Systems, NeurIPS 2020*. NeurIPS.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. In *International Conference on Learning Representations*.

Moosavi-Dezfooli, S.-M.; Fawzi, A.; and Frossard, P. 2016. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2574–2582.

Naseer, M. M.; Ranasinghe, K.; Khan, S. H.; Hayat, M.; Shahbaz Khan, F.; and Yang, M.-H. 2021. Intriguing Properties of Vision Transformers. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*, volume 34, 23296–23308.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Pinto, F.; Torr, P. H.; and K Dokania, P. 2022. An impartial take to the cnn vs transformer robustness contest. In *European Conference on Computer Vision*, 466–480. Springer.

Purushwalkam, S.; and Gupta, A. 2020. Demystifying Contrastive Self-Supervised Learning: Invariances, Augmentations and Dataset Biases. In *Advances in Neural Information Processing Systems*, volume 33.

Qian, R.; Meng, T.; Gong, B.; Yang, M.-H.; Wang, H.; Belongie, S.; and Cui, Y. 2021. Spatiotemporal contrastive video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6964–6974.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763. PMLR.

Wang, F.; and Liu, H. 2021. Understanding the Behaviour of Contrastive Loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2495–2504.

Wang, T.; and Isola, P. 2020. Understanding Contrastive Representation Learning through Alignment and Uniformity on the Hypersphere. In III, H. D.; and Singh, A., eds., *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 9929–9939. PMLR.

Xiao, T.; Wang, X.; Efros, A. A.; and Darrell, T. 2021. What Should Not Be Contrastive in Contrastive Learning. In *International Conference on Learning Representations*.

Xie, C.; Tan, M.; Gong, B.; Wang, J.; Yuille, A. L.; and Le, Q. V. 2020. Adversarial Examples Improve Image Recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.