Can Self-Supervised Representation Learning Methods Withstand Distribution Shifts and Corruptions?

Prakash Chandra Chhipa[1,*], Johan Rodahl Holmgren[1], Kanjar De[1,2,**], Rajkumar Saini[1] and Marcus Liwicki[1]
[1]Machine Learning Group, EISLAB, Luleå Tekniska Universitet, Luleå, Sweden
[2]Video Coding Systems, Fraunhofer Heinrich-Hertz-Institut, Berlin, Germany
[first].[middle].[last]@ltu.se
*corresponding author - prakash.chandra.chhipa@ltu.se

Abstract

Self-supervised representation learning (SSL) in computer vision aims to leverage the inherent structure and relationships within data to learn meaningful representations without explicit human annotation, enabling a holistic understanding of visual scenes. Robustness in vision machine learning ensures reliable and consistent performance, enhancing generalization, adaptability, and resistance to noise, variations, and adversarial attacks. Self-supervised representation learning paradigms, namely contrastive learning, knowledge distillation, mutual information maximization, and clustering, have been considered to have shown advances in invariant learning representations. This work investigates the robustness of learned representations of SSL approaches focusing on distribution shifts and image corruptions in computer vision. Detailed experiments have been conducted to study the robustness of SSL methods on distribution shifts and image corruptions. The empirical analysis demonstrates a clear relationship between the performance of learned representations within SSL paradigms and the severity of distribution shifts and corruptions. Notably, higher levels of shifts and corruptions are found to significantly diminish the robustness of the learned representations. These findings highlight the critical impact of distribution shifts and image corruptions on the performance and resilience of SSL methods, emphasizing the need for effective strategies to mitigate their adverse effects. The study strongly advocates for future research in the field of self-supervised representation learning to prioritize the key aspects of safety and robustness in order to ensure practical applicability. The source code and results are available on GitHub. [1].

[1]https://github.com/prakashchhipa/Robsutness-Evaluation-of-Self-supervised-Methods-Distribution-Shifts-and-Corruptions

[**]*Work performed at Machine Learning Group, EISLAB, Luleå Tekniska Universitet, Luleå, Sweden*

1. Introduction

Safety and robustness are crucial in computer vision as they ensure the accurate and reliable perception of the visual world, enabling applications such as autonomous driving [31], and surveillance systems to make informed and trustworthy decisions, reduce environmental noise [22], ultimately enhancing overall human safety and well-being. In recent years self-supervised representation learning (SSL) methods [18] have garnered interest in computer vision applications. Its current state-of-the-art is prominent even against supervised examples where invariant representation learning has been the core, as stated in [17]. SSL is a well-explored representation learning approach, with many studies on its performance in large datasets such as ImageNet-2012 and also on multi-modality [41]. In addition, SSL has also been well-explored with other learning approaches, including active learning [3], graphs [35], lifelong learning [36], and many more. Recent advances in self-supervised representation learning can be broadly categorized into multiple paradigms, namely contrastive learning [9, 23], Knowledge Distillation [21, 7, 12], Mutual Information Maximization [39, 2], and Clustering [5]. Despite these advancements, the robustness and safety aspects of SSL paradigms have not been extensively explored, which hinders their applicability in real-world use cases. This study is one of the early attempts highlighting the above-stated research gap on a large-scale dataset [25] focusing on the distribution shifts and image corruptions.

Representation learning from self-supervised representation learning paradigms for computer vision can be categorized majorly as (i) Joint Embedding Architecture & Method (JEAM) ([10], [20], [6], [40]), (ii) Prediction Methods ([37], [33], [16]), and loosely (iii) Reconstruction Methods ([30], [19]). Specifically, JEAM can be divided further with each subdivision providing many interesting works; (i) Contrastive Methods (PIRL [32], SimCLR [10], SimCLRv2 [11], MoCo [24]), (ii) Distillation

(BYOL [20], SimSiam [13]), (iii) Quantization (SwAV [6], DeepCluster [4]), and (iv) Information Maximization (Barlow Twins [40], VICReg [1]). Robustness is critical in real-life computer vision applications as there will be a shift in distribution for the deployed models with time. Understanding the behavior of existing models to the distribution shift is a crucial consideration in developing newer, more robust models.

Ericsson et al. [17] explore the impact of different augmentation strategies on the transferability of self-supervised representation learning models to downstream tasks. The authors show that CNNs trained contrastively do learn invariances corresponding to the augmentations used, and specializing CNNs to particular appearance/spatial augmentations can lead to greater corresponding invariances. Furthermore, learning invariances to synthetic transforms does provide a degree of invariance to corresponding real-world transforms. This work establishes the correspondence between synthetic transforms and learning invariances for knowledge transfer limited to [15] without focusing on robustness and distribution shift.

Another significant work by Jiang et al. [28] focuses on improving the robustness of self-supervised pre-training by learning representations that are consistent under both data augmentations and adversarial perturbations. It leverages contrastive learning to enhance adversarial robustness via self-supervised pre-training. They discuss several options to inject adversarial perturbations to reduce adversarial fragility. Through experiments in both supervised fine-tuning and semi-supervised learning settings, they demonstrate that the proposed adversarial contrastive learning can lead to models that are both label-efficient and robust. The paper does not specifically focus on corruption, but rather on improving the model's ability to handle adversarial attacks. This work shows notable improvement in robustness performance but remains limited to a small-scale CIFAR dataset, subject to limited generalizability.

Research is needed to learn invariant SSL representations capable of handling distribution shifts and corruptions; this study provides a ground in this direction by sharing insights into the robustness performance of a large-scale dataset. The identified research gap(s), raises several research questions addressed in later sections. For the detailed investigation, we considered the most popular SSL paradigms, namely contrastive learning, knowledge distillation, mutual information maximization, and clustering. Next, we exhaustively evaluated the corruptions and their severity levels present in ImageNet-C dataset [25] to understand the resilience of each method. Further, compare the robustness performance across multiple metrics, including qualitative analysis. To the best of our knowledge, this is one of the early works in this direction.

Q1: *How do self-supervised representation learning (SSL) paradigms (contrastive learning, knowledge distillation, mutual information maximization, clustering) perform in terms of robustness when exposed to distribution shifts and image corruptions?* A1: Distribution shifts and image corruptions have an effect on the robustness performance of the well-known SSL paradigms. The empirical analysis in this study shows that the error rates (averaged over all distribution shifts and image corruptions) increase with an increase in the severity levels of the distribution shifts and image corruptions. (Figure 1, and Section 3.Q1).

Q2: *To what extent can self-supervised representation learning methods maintain their robustness in the presence of distribution shifts, and what are the factors that limit their ability to do so?* A2: Extensive experiments reveal that SSL methods sustain robustness performance when subjected to lower levels of corruptions, and subsequently, the performance reduces when subjected to higher levels of corruptions. Higher corruptions may lead to massive distribution shifts, which may affect the robustness performance of learned representations. (Figure 2, Table 3, and Section 3.Q2).

Q3: *What is the relationship between the robustness of different SSL paradigms and common categories of corruptions?* A3: Generally, robustness performance decreases for increased severity of corruptions; specifically, the weather group's robustness performance is poorer than that of other groups. (Figure 5, and Section 3.Q3).

Q4: *Do self-supervised representation learning methods deviate from the observed trend of error increase for certain corruptions, and what factors contribute to their robustness in the face of these corruptions?* A4: Yes; a few corruptions, namely, *snow, elastic transform*, and *saturate*, deviate from the observed trend supported by visual quality analysis. (Table 4, and Section 3.Q4).

Q5: *To what extent does the presence of corruptions shift the focus of classifiers from overall representation to specific features?* A5: GradCam [34] analysis reveals that there is a significant shift in the attention maps when the image is subjected to higher levels of corruption. (Figure 3&6, and Section 3.Q5).

Q6: *Do different backbones, such as Convolutional Neural Networks (CNNs) and Transformers, influence the behavior and robustness?* A6. Yes; the self-attention mechanism in transformer, in contrast to CNNs, does not embed any visual inductive bias of spatial locality [27]. (Figure 4, and Section 3.Q6).

2. Methodology

Comparative performance evaluation against robustness is carried out in two steps. In the first step, self-supervised representation learning method(s) are chosen from each potential self-supervised representation learning approach (based on JEAM), including contrastive learning, knowl-

edge distillation, mutual information maximization, and clustering. In the second step, evaluation measures are chosen, indicating quantitative and qualitative comparisons on distribution shift and corrupted data samples from ImageNet-C.

Reason for measuring robustness of learnt representations with corruptions and severity - This study focuses specifically on robustness of representations where domain shifts is simulated in controlled manner through corruption and their varying severity level. Corruptions and perturbations in ImageNet-C [25] are meticulously curated and carefully designed to closely simulate natural phenomena in vision, related to geometric distortions, visual noises, and other explicit factors. Five severity levels further resembles the increased difficulty level, aiding to study robustness at scale. Corruptions across multiple severity levels, thereby altering the original data distribution in a controlled manner [25]. Each corruption severity level shifts the distribution progressively. The corruptions cause variations in texture, color, and spatial coherence, effectively expanding the data manifold towards shift.

2.1. Self-supervised Representation Learning Methods

Methods from different self-supervised representation learning approaches are considered for analysis on the ImageNet-C dataset. The self-supervised representation learning techniques considered for this work are categorized into four main categories based on their methodology.

Contrastive Learning: It is a self-supervised representation learning approach in computer vision and other machine learning domains. The principle behind contrastive learning is to learn valuable representations by encouraging similarity between semantically similar data points while maintaining dissimilarity between unrelated or contrasting data points. In computer vision, this approach helps in learning features and representations from images without relying on labeled data. Instead, it exploits the inherent structure in the data to learn meaningful representations that can be used for various downstream tasks. Specifically, SimCLR method [9] minimizes the temperature-scaled loss function. This contrastive loss penalizes the network when positive pair similarity is low and negative pair similarity is high.

Knowledge Distillation: Distillation-based self-supervision is where student and teacher style encoders are structured and share the learning weights with specific arrangements such as exponential moving averages. Typically, similarity learning is performed by inducing architectural dissimilarity, such as adding a prediction MLP network on only one of the branches. In this work, SimSiam [12], a self-distillation method, and BYOL [21] & DINO (with ResNet encoder) [7] dual encoder style knowledge distillation methods are employed.

Mutual Information Maximization: This principle is used in self-supervised representation learning to learn valuable and meaningful representations from data without explicit labels. The principle is to maximize the mutual information between different views or transformations of the input data, assuming that the learned representations should be invariant or robust to these transformations. Barlow Twins [39], and VICReg [2] are two self-supervised representation learning methods employed for the work to follow the principle of mutual information maximization to learn visual representations by applying redundancy reduction.

Clustering: SwAV [5] combines contrastive learning and clustering-based approaches to learning meaningful and invariant features from images. The main idea behind SwAV is to use a clustering mechanism to enforce consistency between different views of the same image while promoting diversity in the learned representations.

Robustness Evaluation Criteria: The error rate metrics, namely corruption error (CE), mean corruption error (mCE), clean error, average error, and average relative error, were introduced as a standardized measure to benchmark the robustness of machine learning models on Imagenet-C. The two-step evaluation is described by Hendrycks et al. [25]. The same procedure has been followed in this study.

2.2. Dataset and Experimental Setup

ImageNet-C dataset [25] contains 19 types of corruptions with five severity levels, each algorithmically generated. The main objective is to analyze the performance of different self-supervised representation learning methods across these corruptions and severity levels. By conducting detailed experiments, this research aims to gain insights into how self-supervised representations handle various types of corruptions. In this paper, we have performed detailed experiments considering all the corruptions and severity levels to gain a deeper understanding of how different self-supervised representations work on different types of corruptions and present our findings in Section 3.

Table 1. Configuration used (refer [14, 7] for implementation details).

	Barlow Twins	BYOL	SimSiam	SimCLR	DINO	SWaV
Batch Size	2048	4096	256	4096	1024	256
Epochs	300	200	100	200	800	200
Linear-Eval%	71.8	71.8	68.3	66.9	75.3	70.5
Epoch	90	90	90	90	100	100
Batch Size	256	512	512	512	256	256

Experimental details for evaluating the robustness of self-supervised representations are as follows. We extracted the encoder from a ImageNet pre-trained self-supervised representation learning model and added a classification layer at the end of the network. This allows the model to

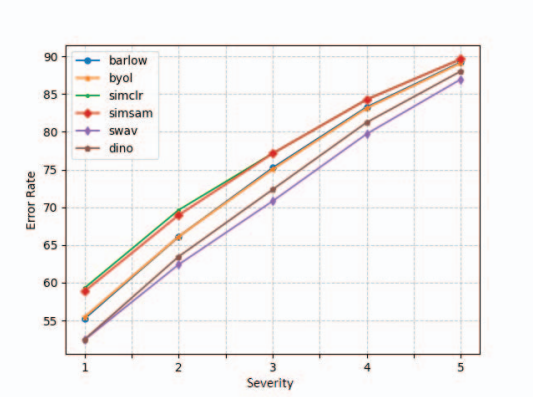

Figure 1. Error rates vs. severity levels across ImageNet-C [25] corruptions.

be fine-tuned on ImageNet 2012 dataset for a classification task. Evaluations is performed on ImageNet-C dataset [25]. For this work, we have considered six of the state-of-the-art SSL algorithms, and the configurations are shown in Table 1. We first initialized the classifier layer randomly and froze all the parameters of the pre-trained encoder. Next, we trained the classifier using the labeled training set. The models used were trained by mmsetup[14] except for DINO, which came from its original repository [7]. ResNet-50 was chosen across all different methods to keep the analysis uniform, and all experiments subsequently were conducted using this architecture. The SSL models were tested on ImageNet-C [25], and mCE [25] is used as a performance measure. The results are shown in Table 2 and 3.

3. Can SSL methods endure shifts in data distribution and image corruptions?

The raised research questions are discussed in this section.

Q1: How do self-supervised representation learning (SSL) paradigms (contrastive learning, knowledge distillation, mutual information maximization, clustering) perform in terms of robustness when exposed to distribution shifts and image corruptions?

The average error rates against all corruptions (per severity level) of all the SSL methods are depicted in Figure 1. The general trend is that SimCLR and SimSam have higher error rates as compared to other methods. While learning has reported good performance previously on ImageNet-C [29], we noticed that SimCLR is not comparably robust against these corruptions. A pattern observed (Figure 1) is that, in general, Knowledge distillation methods seem to outperform contrastive learning. Clustering outperforms other methods indicating robust representations. From Figure 1, one important observation is that for corruptions with lower severity levels, the six SSL methods form three sets where SwaV and DINO perform best, followed by BYOL and Barlow twins; finally, SimCLR and SimSiam have relatively lower performance. However, at the highest severity level, all the methods have similar and high error rates. This is likely because most images in this group are heavily distorted and challenging even for the human visual system to comprehend. From Table 2, we observe that SwaV outperforms all the competing methods in terms of corruption error and mean corruption error; however, DINO has a better robustness performance.

Q2: To what extent can self-supervised representation learning methods maintain their robustness in the presence of distribution shifts, and what are the factors that limit their ability to do so?

Table 3 presents a detailed analysis using mean corruption error mCE for each corruption. Here, we report the average mCE for each corruption in the ImageNet-C dataset. One of the findings is that glass blur significantly impacts the robustness of these models, specifically at higher severity levels. Most of these models have demonstrated good robustness to brightness-based corruptions. As corroborated by Figure 2 for most corruptions, the model robustness suffers with the increase in severity levels.

Q3: What is the relationship between the robustness performance of different SSL paradigms and common categories of corruptions?

As the severity levels of corruptions increase, all self-supervised representation learning (SSL) methods demonstrate a decline in their robustness, as shown in Figure 5. While the *noise* and *blur* groups have a similar trend, whereas *digital* group shows comparatively strong resilience for intermediate severity level. SSL methods are least robust against *weather* group.

Q4: Do self-supervised representation learning methods deviate from the observed trend of error increase for certain corruptions, and what factors contribute to their robustness in the face of these corruptions?

We observed (Figure 2) that for three corruptions, namely, *snow*, *saturate*, and *elastic transform* (last row), there is a deviation from the expected trend; the expected trend is that the error increases with an increase in severity level. However, SSL models are performing low at severity level 2 than at severity level 3. Given the intriguing deviations displayed by (snow, elastic, saturate) from their anticipated behavior, we delved deeper into our inquiry, employing a renowned perceptual measure known as Structural Similarity Index measure (SSIM) [38] to investigate further, as one of the metrics popularly used by image quality researchers for reference image-based quality assessment.

We computed the SSIM between the original image (from ImageNet) and the corresponding corrupted image (from ImageNet-C) for all test images and averaged at each severity level (Table 4); this gives an estimate of the visual

Table 2. Results for each method calculated over the corruption metric [25].

	Barlow Twin	BYOL	SimCLR	SimSam	SWaV	DINO
clean error	28.2	28.2	33.1	31.7	29.5	**24.7**
average error	73.8	73.8	75.99	75.8	**70.5**	71.5
average relative error	74.7	74.6	76.0	76.0	**70.7**	72.9

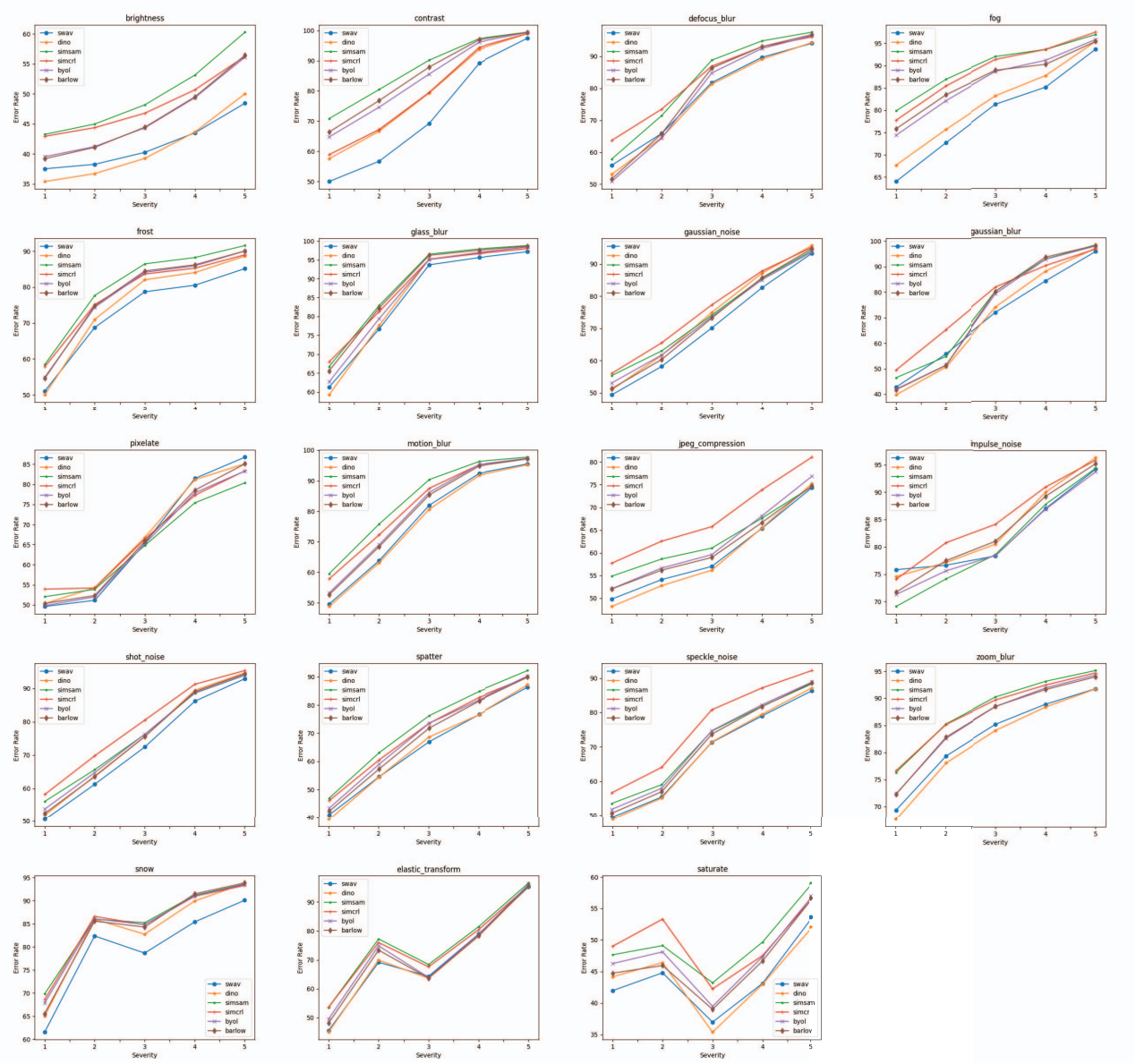

Figure 2. Model performance against specific corruptions by severity. For corruptions, namely, snow, saturate, and elastic (last row), SSL models perform poorly at severity level 2 than at severity level 3.

Table 3. mCE for each corruption type against the baseline. The error rates in each column of corruption types are average values of all severity levels.

		Weather				Noise			Extra				Digital				Blur			
	mCE	Snow	Frost	Fog	Bright	Gauss.	Shot	Impulse	Speckle	Gauss.	Spatter	Saturate	Pixelate	Contrast	Elastic	JPEG	Zoom	Defocus	Motion	Glass
barlow	73.8	84	78	87	46	73	75	83	70	73	69	47	66	85	72	62	86	79	80	88
byol	73.8	85	78	86	46	73	76	81	71	73	70	48	66	84	73	63	86	78	80	86
simclr	76.0	85	78	89	48	76	79	85	76	77	70	50	67	80	75	68	88	83	82	88
simsam	75.8	85	80	90	50	74	76	81	71	74	73	50	65	88	75	63	88	82	84	89
swav	70.5	80	73	79	41	71	73	82	68	70	65	44	67	73	71	60	83	77	77	85
dino	72.9	83	75	82	41	74	75	84	68	70	65	44	68	79	70	60	82	76	76	85
supervised [25]	76.7	78	75	66	57	80	82	83	76	74	76	58	77	71	85	77	80	75	78	89

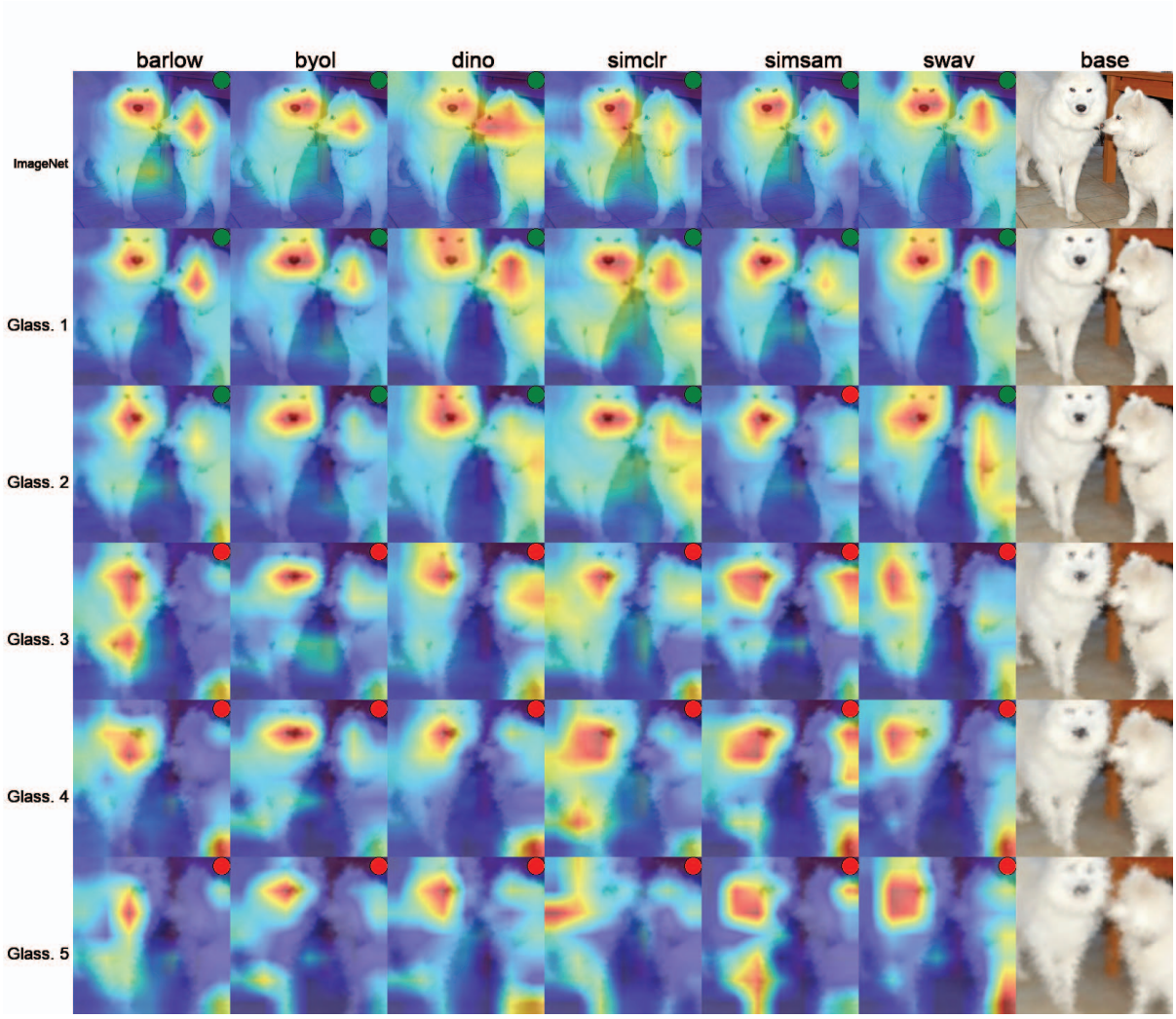

Figure 3. Glass blur on dogs; markers in the images show correct (green) and incorrect (red) classifications. In ImageNet, with many dog breed classes, misclassification doesn't necessarily indicate a bad model if the representation is adequate. In the twin dog example, with low blur severity, both dogs have good activations for all models, suggesting good representations. However, at high blur severity, the model struggles to classify, resulting in distorted activations and difficulty in distinguishing between the dogs, leading to poor results.

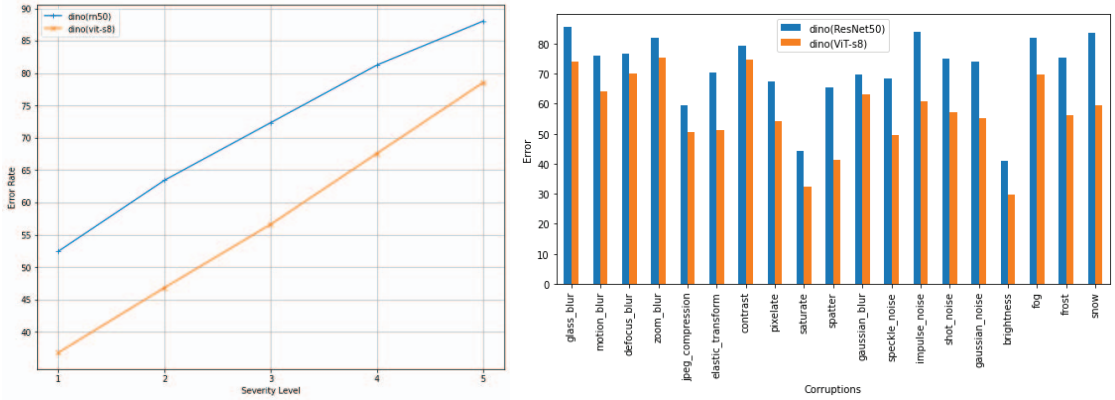

Figure 4. Comparison between different backbones, ResNet50 and ViT-s8 for DINO SSL method over ImageNet-C [25] corruptions. Severity levels (left), corruptions (right).

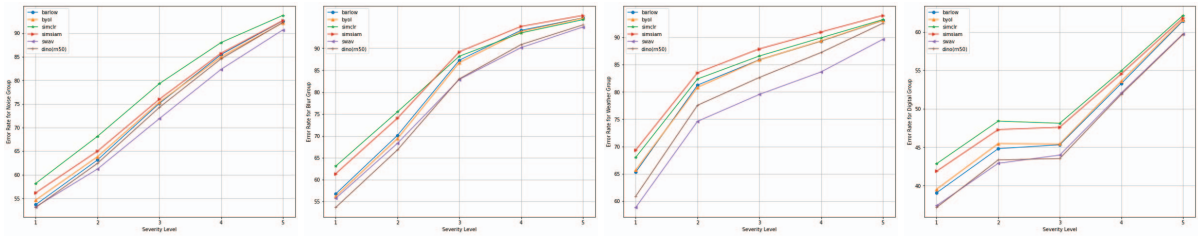

Figure 5. Group-wise comparison. (a) Noise (b) Blur (c) Weather (d) Digital (left to right).

quality.

Snow corruption occludes the object by adding whitish pixels as snowflakes with motion blur. It has more visually challenging images at severity level 2 than other levels; therefore, SSIM at level 2 is lower than the SSIM at other severity levels. Similarly, for *elastic transform*, the SSIM at level 2 is lower than the SSIM at other severity levels. At low severity (levels 1 and 2), the affine transform is more noticeable in some cases, causing artifacts, which can also be seen in figure 6 on its elastic transform. The *saturation* corruptions have very low saturation at low severity levels, causing it to be a grayscale image. This might lead to some classes not being accurately predicted, where color information is crucial. In nutshell, only for snow, elastic and saturate, increased severity level (2 to 4) by increasing respective artifacts, does not reflect increased noise in image examples which mitigates above stated behaviour from all SSL methods.

Table 4. SSIM metric for snow, elastic and saturate-based corruptions.

Severity	Snow	Elastic	Saturate
1	0.218	0.276	0.288
2	0.179	0.237	0.283
3	0.194	0.315	0.273
4	0.186	0.312	0.234
5	0.189	0.305	0.210

Q5: To what extent does the presence of corruptions shift the focus of classifiers from overall representation to specific features?

To gain more insight into how different self-supervised representation learning methods for classification task pick a label, we have used gradcams [34] to compare the different methods qualitatively. Gradcams are used to explain the model's decision as they provide heatmaps on where in the image the model is focusing. In Figure 6, we show the grad cams of an image for all SSL methods under different corruptions of varying severity levels.

The difference among Gradcams gives an understanding of how the model behavior changes in the presence of a particular corruption. From Table 3, we noticed that *glass* blur corruption had caused the highest misclassification for all competing SSL methods; to understand how different methods respond to different severity of *glass* blur, we provide the corresponding gradcams in Figure 3.

Q6: Do different backbones, such as Convolutional Neural Networks (CNNs) and Transformers, influence the behavior and robustness?

There has been analysis [26] on adversarial robustness for transformer and CNN architectures but to specifically analyze the robustness against corruptions and distribution shifts, we chose the most robust SSL method from the previous analysis (i.e., DINO), and compared the backbone ViT-s8 [8] transformer with standard CNN ResNet-50. Undoubtedly, transformer architecture outperformed CNN backbone across the severity levels and also for each image corruption. A detailed trend is shown in Figure 4.

4. Conclusion

The primary objective of this investigation was to conduct an in-depth analysis of diverse paradigms employed in current self-supervised representation learning paradigms, focusing on their robustness characteristics when subjected to varying corruptions present in the ImageNet-C database. The aim was to gain a comprehensive understanding of how these self-supervised representation learning paradigms perform and behave in the face of diverse corruptions, thereby contributing to the advancement of robust representation learning in the computer vision domain. Through empirical analysis, we have presented various analytical trends and demonstrated that self-supervised representation learning methods exhibit decreased robustness as distributional shifts intensify. Notably, our findings indicate that the DINO method employing the distillation approach and the SwAV method utilizing clustering exhibit relatively higher levels of robustness compared to the other methods investigated in this study. While DINO is associated with knowledge distillation, SwAV employs a contrastive assignment quantization approach, indicating their dissimilarity in methodology. These results suggest that multiple SSL methods originating from diverse SSL paradigms display enhanced robustness when evaluated on ImageNet-C. However, it is essential to view these empirical findings as a starting point for further exploration rather than definitive conclusions. The comparative study conducted in this research serves to enhance the comprehension of the com-

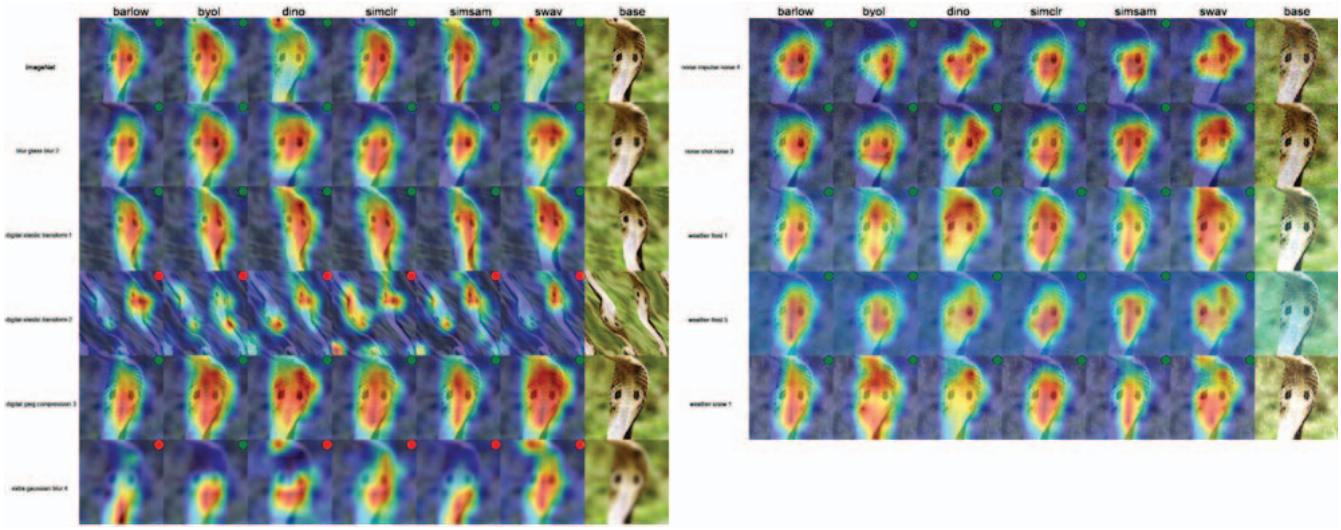

Figure 6. Random Corruptions on a Cobra; the markers in the images show correct and incorrect classifications. Cobra, a reptile with multiple classes in ImageNet, may confuse classifiers. However, cobras are generally distinguishable from other reptiles due to their distinctive neckband. At severity level 2 of elastic transform, there are artifacts causing a distorted doublet, making it a challenging case. Overall, models perform well with good representation. The classifier in these examples shows a bias toward the neckband, while the original un-cropped image emphasizes the edges for classification by DINO and SwaV. However, due to corruptions, the focus shifts more towards the neckband.

puter vision community regarding the strengths and limitations of various self-supervised representation learning approaches. Furthermore, it facilitates researchers in developing robust representations in future endeavors. A significant finding from our analysis is that the SwaV method, which employs a clustering approach, exhibits higher robustness compared to popular methods such as SimCLR and Barlow Twins. This result offers valuable insights for future research directions aimed at further improving self-supervised representation learning methodologies. Considering the findings of this study, it becomes imperative to address the challenges associated with the performance degradation of self-supervised representation learning methods under distribution shifts and image corruptions. By prioritizing safety and robustness, researchers can contribute to the development of more reliable and trustworthy self-supervised representation learning techniques that can effectively handle real-world scenarios and enhance the practical utility of these methods. In this work, we dedicate to the methodical revelation of empirical evidence, rather than hypothesizing. Our endeavor remains steadfast in illuminating numerous enigmas through a rigorous examination. We firmly hold the conviction that this pioneering work shall pave the way for future inquiries, enabling the formulation and evaluation of cogent hypotheses.

References

[1] Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. arXiv preprint arXiv:2105.04906, 2021.

[2] Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. In ICLR, 2022.

[3] Razvan Caramalau, Binod Bhattarai, D Stoyanov, and Tae-Kyun Kim. Mobyv2: Self-supervised active learning for image classification. In The 33rd British Machine Vision Conference (BMVC). BMVA, 2022.

[4] 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), pages 132–149, 2018.

[5] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. Advances in neural information processing systems, 2020.

[6] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. arXiv preprint arXiv:2006.09882, 2020.

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In Proceedings of the International Conference on Computer Vision (ICCV), 2021.

[8] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In Proceedings of the IEEE/CVF international conference on computer vision, pages 9650–9660, 2021.

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020.

[10] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In International conference on machine learning, pages 1597–1607. PMLR, 2020.

[11] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. Advances in Neural Information Processing Systems, 33:22243–22255, 2020.

[12] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In CVPR, 2021.

[13] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 15750–15758, 2021.

[14] MMSelfSup Contributors. MMSelfSup: Openmmlab self-supervised learning toolbox and benchmark. https://github.com/open-mmlab/mmselfsup, 2021.

[15] 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, pages 248–255. Ieee, 2009.

[16] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In Proceedings of the IEEE international conference on computer vision, pages 1422–1430, 2015.

[17] Linus Ericsson, Henry Gouk, and Timothy M Hospedales. Why do self-supervised models transfer? on the impact of invariance on downstream tasks. In The 33rd British Machine Vision Conference, 2022, page 509. BMVA Press, 2022.

[18] Linus Ericsson, Henry Gouk, Chen Change Loy, and Timothy M Hospedales. Self-supervised representation learning: Introduction, advances, and challenges. IEEE Signal Processing Magazine, 39(3):42–62, 2022.

[19] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. Advances in neural information processing systems, 27, 2014.

[20] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. Advances in neural information processing systems, 33:21271–21284, 2020.

[21] 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, et al. Bootstrap your own latent: A new approach to self-supervised learning. In NeurIPS, 2020.

[22] Jianfeng He, Bei Xiao, Xuchao Zhang, Shuo Lei, Shuhui Wang, and Chang-Tien Lu. Reducing noise pixels and metric bias in semantic inpainting on segmentation map. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 1876–1885, 2021.

[23] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning, 2020.

[24] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 9729–9738, 2020.

[25] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. Proceedings of the International Conference on Learning Representations, 2019.

[26] Jaehyuk Heo, Seungwan Seo, and Pilsung Kang. Exploring the differences in adversarial robustness between vit-and cnn-based models using novel metrics. Computer Vision and Image Understanding, page 103800, 2023.

[27] Samy Jelassi, Michael Sander, and Yuanzhi Li. Vision transformers provably learn spatial structure. Advances in Neural Information Processing Systems, 35:37822–37836, 2022.

[28] Ziyu Jiang, Tianlong Chen, Ting Chen, and Zhangyang Wang. Robust pre-training by adversarial contrastive learning. Advances in neural information processing systems, 33:16199–16210, 2020.

[29] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning, 2021.

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

[31] Jiale Li, Hang Dai, and Yong Ding. Self-distillation for robust lidar semantic segmentation in autonomous driving. In Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVIII, pages 659–676. Springer, 2022.

[32] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6707–6717, 2020.

[33] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI, pages 69–84. Springer, 2016.

[34] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE international conference on computer vision, pages 618–626, 2017.

[35] Amir Shirian, Krishna Somandepalli, and Tanaya Guha. Self-supervised graphs for audio representation learning with limited labeled data. IEEE Journal of Selected Topics in Signal Processing, 16(6):1391–1401, 2022.

[36] Mamatha Thota, Dewei Yi, and Georgios Leontidis. Lleda–lifelong self-supervised domain adaptation. arXiv preprint arXiv:2211.09027, 2022.

[37] Bastiaan S Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant cnns for digital pathology. In International Conference on Medical image computing and computer-assisted intervention, pages 210–218. Springer, 2018.

[38] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing, 13(4):600–612, 2004.

[39] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In International Conference on Machine Learning, 2021.

[40] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In International Conference on Machine Learning, pages 12310–12320. PMLR, 2021.

[41] Yongshuo Zong, Oisin Mac Aodha, and Timothy Hospedales. Self-supervised multimodal learning: A survey. arXiv preprint arXiv:2304.01008, 2023.