

# Pre-train to Gain: Robust Learning Without Clean Labels

David Szczecina,

Nicholas Pellegrino, & Paul Fieguth

Vision and Image Processing Group, Systems Design Engineering, University of Waterloo

{david.szczecina, npellegr, pfieguth}@uwaterloo.ca

## Abstract

*Training deep networks with noisy labels leads to poor generalization and degraded accuracy due to overfitting to label noise. Existing approaches for learning with noisy labels often rely on the availability of a clean subset of data. By pre-training a feature extractor backbone without labels using self-supervised learning (SSL), followed by standard supervised training on the noisy dataset, we can train a more noise robust model without requiring a subset with clean labels. We evaluate the use of SimCLR and Barlow Twins as SSL methods on CIFAR-10 and CIFAR-100 under synthetic and real world noise. Across all noise rates, self-supervised pre-training consistently improves classification accuracy and enhances downstream label-error detection (F1 and Balanced Accuracy). The performance gap widens as the noise rate increases, demonstrating improved robustness. Notably, our approach achieves comparable results to ImageNet pre-trained models at low noise levels, while substantially outperforming them under high noise conditions.*

## 1. Introduction

Deep neural networks have achieved remarkable success across a wide range of supervised learning tasks [6, 12]. However, their performance is heavily dependent on the quality of labeled data. In real-world settings, datasets often contain noisy labels due to human annotation errors, automated labeling processes, or ambiguous data, where  $\eta \in [0, 1]$  denotes the fraction of labels that are incorrect [15, 27]. Training directly on such noisy labels typically leads to poor generalization, as deep networks are prone to memorizing the noise [19, 20, 26], especially in the absence of explicit noise-handling mechanisms.

A common strategy to mitigate this issue is to identify and correct noisy labels using robust loss functions or noise-aware training strategies [5, 13, 22, 28]. Many of these methods, however, assume access to a small subset of clean labels ( $\eta = 0$ ) or require prior knowledge about the noise distribution [5, 13, 14], assumptions that are often unrealistic in practical scenarios. Self-supervised learning

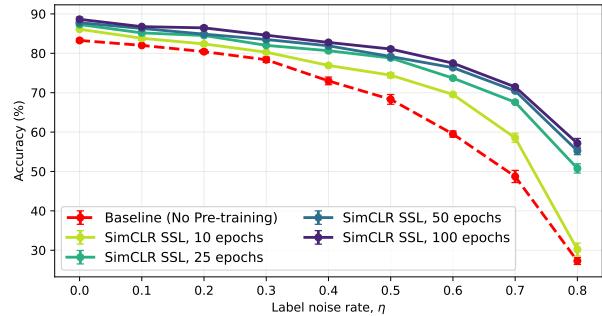


Figure 1. Classification accuracy across varying label noise rates, for increasing durations of self-supervised pre-training on CIFAR-10 using SimCLR. Mean over 5 seeds with standard error is graphed. As the number of SSL pre-training epochs increases, downstream accuracy consistently improves across all corruption levels.

(SSL) offers a promising alternative by allowing feature representations to be learned without using labels at all [1, 2, 4, 8, 21, 25]. SSL methods such as SimCLR [1] and Barlow Twins [25] leverage data augmentations and contrastive or redundancy-reduction objectives to train encoders that capture semantically meaningful representations. By using SSL, feature extractors can be created without considering labels, thereby avoiding any negative influences of mislabelled data.

In this work, we investigate whether self-supervised pre-training can enhance a model’s noise robustness when training on noisy-labelled datasets, without relying on any clean subset of data. We perform experiments on CIFAR-10 and CIFAR-100 [11], both with synthetically introduced label noise, and with real-world noisy labels from the CIFAR-10N and CIFAR-100N datasets [23]. We compare a standard supervised training pipeline against a two-stage approach where a ResNet18 [7] backbone is first pre-trained using self-supervised learning and then fine-tuned with supervised training on the same noisy dataset.

As seen by results in Figure 1, self-supervised pre-training consistently improves model robustness. Specifically, we observe higher test accuracy, evaluated on clean test sets, and improved label error detection capability when predicting on noisy test sets.

We further compare our SSL pre-training scheme against the use of ImageNet pre-trained weights for model initialization [3], finding comparable performance at low rates of label noise, but the effectiveness of ImageNet weights diminishes with higher noise rates. In contrast, self-supervised pre-training continues to offer substantial performance improvements under severe label noise.

This study demonstrates that self-supervised pre-training on the target dataset can serve as a simple yet powerful method for improving models robustness to label errors when training on noisy datasets.

## 2. Background

**Learning with Noisy Labels.** Supervised learning assumes access to correctly labelled data, but in real-world scenarios, label noise is common [16]. Training deep neural networks directly on noisy data often results in memorization of incorrect labels, leading to poor generalization and overfitting [19, 20, 26].

Numerous strategies have been proposed to address training with label noise, including:

- Robust loss functions that reduce the model’s sensitivity to incorrect labels [18, 24, 27].
- Sample re-weighting to prioritize clean samples [5, 9].
- Semi-supervised and weakly supervised approaches that leverage a clean subset to guide training [13, 14].

A recurring limitation across many of these approaches is the reliance on a small subset of data with clean labels, which may not be feasible in large-scale or low-resource settings [16].

**Self-Supervised Learning (SSL).** Self-supervised learning has emerged as a powerful method for representation learning without requiring any labels [1, 4, 25]. In computer vision, SSL methods typically use data augmentations and auxiliary pretext tasks to learn meaningful image embeddings. Two widely used approaches are SimCLR, a contrastive learning method [1], and Barlow Twins, which is a redundancy reduction method [25]. These methods have been shown to learn generalizable and semantically rich features, even without labels. Importantly, when used for pre-training, these SSL representations can significantly improve downstream performance by providing better model initialization.

**Label Error Detection.** Identifying mislabelled instances is a critical step in improving training with noisy datasets. Frameworks for detecting label errors typically require well-trained and well-generalized models to perform effectively. Several techniques have been proposed for this purpose, ranging from training dynamics based methods [19] to confidence-based filtering. Confident Learning [15] is a prominent framework that estimates the probability that

each example’s given label is incorrect. It does so using a combination of predicted probabilities and the confusion matrix estimated from the model’s outputs.

**Related Works.** Contrast to Divide (C2D) [28] identified a key limitation in learning with noisy labels, the “warm-up obstacle”, where models quickly begin memorizing incorrect labels during initial supervised training. C2D addresses this by introducing a two-stage framework: a self-supervised contrastive pre-training phase followed by integration with existing learning-with-noisy-labels (LNL) algorithms such as DivideMix [13] or ELR+ [14]. This pre-training step provides noise-invariant features that improve label separation and classification accuracy.

In contrast to C2D, our work isolates the effect of self-supervised pre-training itself, without coupling it to any LNL algorithm or specialized loss. Rather than improving a specific warm-up mechanism, we show that stand-alone SSL pre-training on the same noisy dataset consistently enhances downstream performance for both classification accuracy and the models ability to detect label-errors.

## 3. Method

To evaluate the effectiveness of self-supervised pre-training in enhancing robustness to label noise, we compare a baseline standard supervised training method against a two-stage training pipeline that incorporates SSL pre-training.

- **Baseline Supervised Method:** The model is trained from scratch on the noisy dataset for 10 epochs.
- **Self-Supervised + Supervised Fine-tuning:** An identical model is first pre-trained using an SSL method, for a specified number of epochs, before being fine-tuned via 10-epoch supervised training on the noisy dataset.

**Self-Supervised Methods.** To ensure that our findings are not inherent to a specific SSL objective, we evaluate two complementary SSL methods: SimCLR (contrastive) [1] and Barlow Twins (non-contrastive) [25]. This allows us to test whether robustness to label noise persists across fundamentally different families of SSL methods. SimCLR is implemented following its standard setup from Chen et al. [1], with random cropping, colour jitter, and Gaussian blur augmentations, while Barlow Twins is applied using its default configuration from Zbontar et al. [25]. Both methods train on the entire noisy train dataset, without any label information.

**Datasets and Artificial Corruption.** Synthetic uniform label noise is injected into standard benchmark datasets following the same method used in Pellegrino et al. [17]. For a given corruption rate  $\eta \in [0, 1]$ , a fraction  $\eta$  of the training labels are randomly flipped. Two benchmark image classification datasets are used, CIFAR-10 and CIFAR-100, under the assumption that their existing label error rates [16]

are negligible compared to the induced corruption,  $\eta$ . Real-world noise is explored using the CIFAR-N datasets [23], wherein CIFAR-100N has noise rate  $\eta \approx 0.4$ , and CIFAR-10N has two sets of noisy labels, with  $\eta \approx 0.1$  and  $\eta \approx 0.4$ .

**Label Error Detection.** To assess how well the trained models can identify incorrect labels, we use Confident Learning, a framework that estimates the probability of label error based on the model’s predicted probability outputs [15]. We evaluate the model on the corrupted test set, generating predicted probabilities for Confident Learning, and compute the Balanced Accuracy and F1 Score for correctly identified label errors in the test set. These metrics serve as indicators of the model’s robustness to label errors, represented by the downstream ability to distinguish between correctly and incorrectly labelled data points.

**Model Training.** To ensure fairness, models are trained only for the number of epochs at which point overfitting begins on the original uncorrupted datasets. A precursor experiment was performed for both uncorrupted datasets whereby the crossover point between training and testing loss is measured. In both cases, overfitting was found to occur starting at roughly 10 epochs. In all cases, a ResNet-18 model architecture and the Adam optimizer [10] is used.

## 4. Results

**Main experiment.** In Table 1, we compare the performance of supervised only training against models with SSL pre-training done using either SimCLR or Barlow Twins. Pre-training was performed for 100 epochs, and all results are averaged over 5 seeds. The classification accuracy on a clean test set was always higher with the SSL pre-trained methods, and significantly increased performance on datasets with higher amounts of corruption. On CIFAR-10, SimCLR SSL boosts clean test accuracy by 5–30%, while on CIFAR-100 the improvement ranges from 6–20% when using Barlow Twins. On a noisy test set, Confident Learning is used to detect label errors, for which the F1 and Balanced Accuracy scores are reported. SSL pre-trained

methods improved the F1 score and Balanced Accuracy by 4–7 percentage points on average for all levels of corruption on both datasets.

**Real-World Noise.** To evaluate performance under real-world, non-uniform label errors, we additionally test on the CIFAR-10N and CIFAR-100N datasets, which contain human annotation noise. This experiment mirrors the main setup, evaluating the baseline and SimCLR pre-trained models. Results averaged over five seeds are reported in Table 2. Based on optimal pre-train durations from Figure 2, we pre-train for 50 epochs on CIFAR-10 and 25 epochs on CIFAR-100. Pre-training increased performance on all datasets, with a 5–8% improvement in classification accuracy being measured.

Table 2. Performance comparison on real-world noisy datasets CIFAR-10N and CIFAR-100N. Accuracy, F1-score, and Balanced Accuracy are shown for baseline and SimCLR-pretrained models.

Dataset	Accuracy (%)		F1		BA	
	Baseline	SimCLR	Baseline	SimCLR	Baseline	SimCLR
CIFAR-10N, $\eta \approx 0.1$	81.14	<b>85.84</b>	53.70	<b>59.69</b>	74.48	<b>76.40</b>
CIFAR-10N, $\eta \approx 0.4$	72.73	<b>78.71</b>	77.46	<b>80.74</b>	81.18	<b>83.77</b>
CIFAR-100N, $\eta \approx 0.4$	43.91	<b>47.60</b>	64.20	<b>65.89</b>	69.72	<b>71.23</b>

**Pre-train Duration variation.** We examine how the duration of SSL pre-training influences downstream robustness by evaluating models pretrained for different numbers of epochs. As shown in Figure 2, at just 10 epochs of pre-training using SimCLR there are noticeable improvements, but the gains begin to plateau after approximately 50 epochs. This trend is consistent across both CIFAR-10 and CIFAR-100, and across both SSL methods evaluated. Comparable patterns are observed when examining classification accuracy, as seen in Figure 1, and Balanced Accuracy, indicating that early-stage pre-training provides most of the benefit, with diminishing returns beyond moderate training durations. During this early stage, the encoder learns clean, noise-agnostic features before any corrupted labels can bias the representation.

Table 1. Comparison of classification accuracy % (Acc), F1-score (F1), and Balanced Accuracy (BA) under varying noise rates  $\eta$  on CIFAR-10 and CIFAR-100. Results are averaged over five seeds, and best scores are bolded. Self-supervised pretraining for 100 epochs with either SimCLR or Barlow Twins consistently improves performance under label noise.

Dataset	Method	$\eta = 0.0$		$\eta = 0.2$			$\eta = 0.4$			$\eta = 0.6$			$\eta = 0.8$		
		Acc		Acc	F1	BA									
CIFAR-10	Baseline	83.27		80.42	0.80	0.91	73.00	0.85	0.87	59.50	0.83	0.79	27.28	0.73	0.61
	SimCLR	<b>88.63</b>		<b>86.46</b>	<b>0.84</b>	<b>0.93</b>	<b>82.76</b>	<b>0.90</b>	<b>0.92</b>	<b>77.50</b>	<b>0.91</b>	<b>0.88</b>	<b>57.16</b>	<b>0.76</b>	<b>0.73</b>
	Barlow Twins	87.95		85.09	0.83	0.92	81.74	0.89	0.91	74.89	0.90	0.87	51.12	0.75	0.71
CIFAR-100	Baseline	56.21		50.03	0.57	0.78	41.40	0.72	0.75	27.42	0.77	0.69	9.74	0.68	0.61
	SimCLR	59.41		55.17	0.59	0.79	50.43	0.75	0.78	41.79	0.82	0.75	22.95	0.77	0.67
	Barlow Twins	<b>62.76</b>		<b>58.07</b>	<b>0.63</b>	<b>0.81</b>	<b>53.05</b>	<b>0.76</b>	<b>0.80</b>	<b>45.31</b>	<b>0.83</b>	<b>0.77</b>	<b>30.89</b>	<b>0.82</b>	<b>0.71</b>

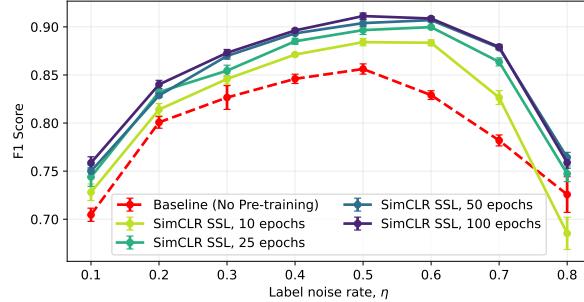


Figure 2. Effect of SimCLR SSL pre-training duration on label error detection (F1-score) on CIFAR-10. Mean over 5 seeds is graphed. Performance improves rapidly with additional pre-training, with gains plateauing after approximately 50 epochs.

**ImageNet pre-trained weights.** Often, in practice, models are initialized with ImageNet pre-trained weights, which are widely regarded as strong general-purpose feature extractors [12]. To assess the real-world relevance of our approach, we compare our SSL pre-training scheme directly against ImageNet initialization in Table 3. At low corruption rates, both initialization strategies yield similar performance. However, as label noise increases, ImageNet pre-training offers little to no advantage over training from scratch, while SSL pre-training continues to provide substantial improvements. This highlights the benefit of domain-aligned self-supervised pre-training when working with noisy datasets.

Table 3. Comparison of self-supervised pre-training and ImageNet initialization under increasing label noise  $\eta$  on CIFAR-100.

Method	Accuracy (%)				
	$\eta = 0.0$	$\eta = 0.2$	$\eta = 0.4$	$\eta = 0.6$	$\eta = 0.8$
Baseline	83.27	80.42	73.00	59.50	27.28
ImageNet Pretrained	<b>89.43</b>	<b>87.25</b>	81.88	74.02	33.67
SSL Pretrained	88.63	86.46	<b>82.76</b>	<b>77.50</b>	<b>57.16</b>

**Increased Supervised Training.** To test whether the benefits of SSL pre-training persist when models are allowed substantially more supervised training, we extend the fine-tuning stage from 10 to 100 epochs. As shown in Figure 3, averaged over 5 seeds, models trained from scratch take longer to reach peak performance and then rapidly overfit to the corrupted labels, causing a sharp decline in test accuracy and a pronounced rise in training loss.

In contrast, SSL pretrained models remain far more stable throughout training: they achieve higher accuracy, overfit more slowly, and exhibit a much weaker increase in loss during the overfitting phase. This behaviour indicates that the noise-resilient representations learned during pre-training continue to shield the model from memorizing corrupted labels, and even last for prolonged supervised training.

These results highlight that the gains from SSL pre-training are not merely from increased total training duration, but from improved and more robust feature representations.

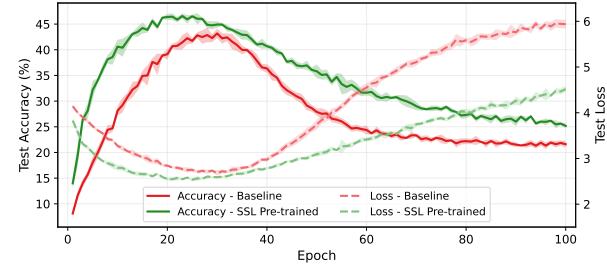


Figure 3. Comparison of Baseline and SSL-pretrained models over 100 supervised epochs on Cifar-100 with  $\eta = 0.6$ . SSL pre-training (using SimCLR for 25 epochs) yields higher accuracy, slower overfitting, and reduced loss escalation under label noise.

## 5. Discussion

Our results show that self-supervised pre-training on the same noisy dataset used for downstream training substantially improves robustness to label errors. Because SSL does not use labels, the encoder learns structure directly from the data rather than from corrupted annotations. Fine-tuning then begins from a feature space that already exhibits clean class separation, reducing the likelihood that the model memorizes incorrect labels.

Across all corruption levels, and especially under heavy noise ( $\eta \geq 0.6$ ), self-supervised pre-training consistently improves both test accuracy and label-error detection compared to training from scratch. While ImageNet pre-trained backbones perform comparably to SSL-pretrained models at low noise, their benefits quickly diminish as corruption increases. In contrast, SSL pre-training retains strong performance even under severe noise, indicating that features learned directly from the target domain are more robust to label corruption than features obtained from external datasets. We also find that most of the benefits of SSL emerge early, with a minimal amount training capturing most of the robustness gains, after which improvements plateau. The noise-resilient representations additionally last for prolonged supervised training, continuing to limit overfitting to corrupted labels. This makes the approach computationally practical, and cost-effective for noisy real-world datasets.

Overall, our results underscore self-supervised pre-training as a simple, scalable, and broadly applicable strategy for noise-robust learning, that enhances downstream performance without modifying the supervised training pipeline or introducing additional assumptions.

## 6. Conclusion

This work demonstrates that self-supervised pre-training on a noisy target dataset is an effective and practical way to improve robustness to label errors. Future work will explore scaling to larger datasets, and the integration of modern SSL techniques which are designed for Vision Transformers.

## References

- [1] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. *CoRR*, abs/2002.05709, 2020. [1](#), [2](#)
- [2] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. *CoRR*, abs/2011.10566, 2020. [1](#)
- [3] 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. [2](#)
- [4] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. *CoRR*, abs/2006.07733, 2020. [1](#), [2](#)
- [5] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor W. Tsang, and Masashi Sugiyama. Co-sampling: Training robust networks for extremely noisy supervision. *CoRR*, abs/1804.06872, 2018. [1](#), [2](#)
- [6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. [1](#)
- [7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. [1](#)
- [8] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. *CoRR*, abs/1911.05722, 2019. [1](#)
- [9] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. MentorNet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *International conference on machine learning*, pages 2304–2313. PMLR, 2018. [2](#)
- [10] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. [3](#)
- [11] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical Report 0, University of Toronto, Toronto, Ontario, 2009. [1](#)
- [12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 1*, page 1097–1105. Curran Associates Inc., 2012. [1](#), [4](#)
- [13] Junnan Li, Richard Socher, and Steven C. H. Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. *CoRR*, abs/2002.07394, 2020. [1](#), [2](#)
- [14] Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels. *CoRR*, abs/2007.00151, 2020. [1](#), [2](#)
- [15] Curtis Northcutt, Lu Jiang, and Isaac Chuang. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 70:1373–1411, 2021. [1](#), [2](#), [3](#)
- [16] Curtis G Northcutt, Anish Athalye, and Jonas Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks. *Advances in Neural Information Processing Systems*, 2021. [2](#)
- [17] Nicholas Pellegrino, Nolen Zhao, and Paul Fieguth. The effects of label errors in training data on model performance and overfitting. *Journal of Computational Vision and Imaging Systems*, 9(1):26–29, 2023. [2](#)
- [18] Nicholas Pellegrino, David Szczecina, and Paul Fieguth. Loss functions robust to the presence of label errors. *Journal of Computational Vision and Imaging Systems*, 10(1):24–29, 2024. [2](#)
- [19] Geoff Pleiss, Tianyi Zhang, Ethan Elenberg, and Kilian Q Weinberger. Identifying mislabeled data using the area under the margin ranking. *Advances in Neural Information Processing Systems*, 33:17044–17056, 2020. [1](#), [2](#)
- [20] Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. *IEEE transactions on neural networks and learning systems*, 34(11):8135–8153, 2022. [1](#), [2](#)
- [21] Zahra Vaseqi, Ibtihel Amara, and Samrudhdhi Rangrej. Label noise resiliency with self-supervised representations. In *NeurIPS 2021 Workshop on Self-Supervised Learning: Theory and Practice (SSL-NeurIPS)*, 2021. [1](#)
- [22] Hongxin Wei, Lei Feng, Xiangyu Chen, and Bo An. Combating noisy labels by agreement: A joint training method with co-regularization. *CoRR*, abs/2003.02752, 2020. [1](#)
- [23] Jiaheng Wei, Zhaowei Zhu, Hao Cheng, Tongliang Liu, Gang Niu, and Yang Liu. Learning with noisy labels revisited: A study using real-world human annotations. In *International Conference on Learning Representations*, 2022. [1](#), [3](#)
- [24] Xichen Ye, Xiaoqiang Li, Tong Liu, Yan Sun, Weiqin Tong, et al. Active negative loss functions for learning with noisy labels. *Advances in Neural Information Processing Systems*, 36:6917–6940, 2023. [2](#)
- [25] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. *CoRR*, abs/2103.03230, 2021. [1](#), [2](#)
- [26] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016. [1](#), [2](#)
- [27] Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31, 2018. [1](#), [2](#)
- [28] Evgenii Zheltonozhskii, Chaim Baskin, Avi Mendelson, Alex M. Bronstein, and Or Litany. Contrast to divide: Self-supervised pre-training for learning with noisy labels. *CoRR*, abs/2103.13646, 2021. [1](#), [2](#)