No Data Augmentation?
Alternative Regularizations for Effective Training on Small Datasets

Anonymous ICCV submission
Paper ID ****

Abstract

Solving image classification tasks given small training datasets remains an open challenge for modern computer vision. Aggressive data augmentation and generative models are among the most straightforward approaches to overcoming the lack of data. However, the first fails to be agnostic to varying image domains, while the latter requires additional compute and careful design. In this work, we study alternative regularization strategies to push the limits of supervised learning on small image classification datasets. In particular, along with the model size and training schedule scaling, we employ a heuristic to select (semi) optimal learning rate and weight decay couples via the norm of model parameters. By training on only 1% of the original CIFAR-10 training set (i.e., 50 images per class) and testing on ciFAIR-10, a variant of the original CIFAR without duplicated images, we reach a test accuracy of 66.5%, on par with the best state-of-the-art methods.

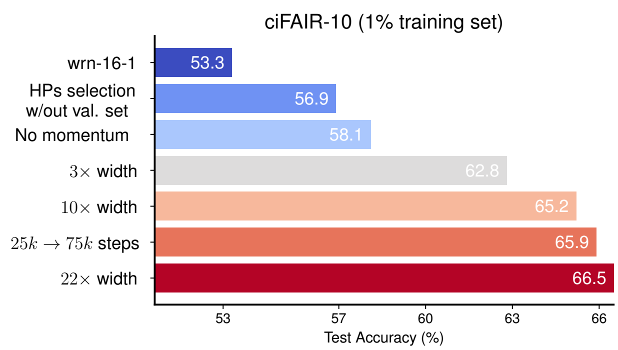

Figure 1: **Overview.** We examine the impact of each additional modification in our training setup beyond the baseline WRN-16-1 model, which has been tuned for optimal learning rate and weight decay on a subset of the small training set.

1. Introduction

In recent years, significant progress has been made in computer vision through large-scale pretraining on extensive datasets [55, 52]. However, improving the data efficiency of deep neural networks and enabling successful training on significantly smaller datasets, ranging from a few tens to hundreds of images per class, remains an ongoing area of research. Better sample efficiency and generalization would greatly benefit domains where the high cost and limited accessibility of data collection and annotation are critical barriers (e.g., the medical domain). The community has recently increased its focus toward studying limited-sample problems with deep learning through the organization of dedicated workshops and challenges [12, 37, 13]. Furthermore, recent work has compared methods tailored explicitly for image classification with small datasets and established a dedicated benchmark [9, 10]. A notable result of the latter analysis regards the importance of hyper-parameter tuning, particularly weight decay, which plays a significant role in the generalization ability of networks and has often been overlooked in previous works. More in detail, a tuned vanilla cross-entropy classifier favorably compared against most of the evaluated data-efficient methods, powered by sophisticated techniques (e.g., [7, 45]) and inductive biases (e.g., [47, 31]).

Classifiers augmented with aggressive data augmentation methods (e.g., AutoAugment [15]) or generative models have recently scored the best results on multiple data-efficient image classification benchmarks [2, 51]. While it is expected that additional data synthesis helps generalization, this family of approaches still presents challenges. For instance, when the image domain differs from popular object-centric natural images, hand-crafted augmentations can also introduce strong biases that may be detrimental and not transfer to different data distributions. While generative models overcome such an issue, they require sophisticated design, careful engineering, and multi-stage training [73, 2, 51].

A large body of work has investigated alternative regularization strategies for deep neural networks, such as the scaling of model size [23, 69, 57], training length [27], and the effect of L2 regularization [60]. Larger models tend to generalize better on large datasets and to follow the *deep double descent* phenomenon, named after the peculiar shape of the generalization-error over the model-size curve, observed empirically in multiple works [6, 1]. However, it remains challenging to train networks on very few samples properly [9, 10]. Over-parametrization provides the model with more possible solutions [14], but the network is keen to overfit without proper regularization. Additionally, since most of the state-of-the-art architectures are scale-invariant [60, 26, 40], their optimization dynamics are not yet fully understood and strongly impacted by weight decay, which plays an important role.

In this work, we investigate in detail the impact of optimization-related hyper-parameters (HPs) (i.e., learning rate, weight decay, and momentum), model size (in particular width), and training schedule length on the popular ciFAIR-10 small-data benchmark, which comprises 1% of the original training set of CIFAR-10 and testing set without duplicated images [5]. Based on our empirical analysis, we devise a simple scheme to maximize the accuracy of a vanilla cross-entropy classifier by making it as data-efficient as state-of-the-art methods powered by strong data augmentation methods [19, 51]. As visible in Fig. 1, we start from a baseline Wide ResNet-16-1 (WRN-16-1) [69], tuned on the small validation set, which scores 53.3% on the test set, and reach a strong 66.5% accuracy with WRN-16-22.

In summary, this paper builds a robust and easy-to-implement baseline for training efficiently vanilla cross-entropy classifiers on small datasets. Furthermore, it provides insights regarding the impact of HPs, model scale, and training length. We demonstrate that aggressive data augmentation is not the only way to reach the best performance in scenarios with limited data. We hope that our empirical analysis could be helpful for practitioners and researchers involved in deploying and searching for more data-efficient image classifiers.

2. Related Work

Impact of scaling model size and training length. Several studies have explored the effect of model scaling on performance. For instance, convolutional networks can be scaled by depth [23], width [69], or the combination of the two along with the input resolution [57]. Other works studied the generalization of networks across data and model scaling [25, 53], with some focusing on small data regimes [8, 11]. The relationship between generalization error and model size, with the empirical finding of the *double descent* phenomenon, has been widely investigated [46, 6, 44]. Although models of different sizes reach the same training errors, larger models tend to have smaller test errors [1]. While still under discussion in current research, possible explanations include that large models are more biased towards better minima [17, 16] or explore more features [14]. Finally, additional training iterations benefit generalization [27, 22], and seem to generate a similar *double descent* behavior but related to the length of training [44, 50].

Scale-invariant networks. Normalization layers (e.g., Batch Normalization (BN) [28]) make modern neural networks almost fully scale-invariant. In other words, their output activations, and consequently, the loss function, does not change if the weights undergo scaling, implying that weight decay does not limit the model capacity as previously believed [60]. The training dynamics of Stochastic Gradient Descent (SGD) and variants have been widely investigated and are still under discussion from both an empirical and theoretical perspective [26, 70, 40, 62, 35]. The parameters' norm strongly impacts the effective learning rate, the actual step which a scale-invariant network would take if optimized over the unit sphere [62, 35]. Recent work has practically studied predicting and scheduling optimal HPs by exploiting SGD symmetries as data scales [67, 68].

Image classification with small datasets. Learning from a small sample is an actual challenge for deep learning. and shares the goal of deploying data-efficient classifiers with other popular research areas, such as transfer learning [49, 36], domain adaptation [63], and few-shot learning [64]. However, such research domains assume access to a generally extensive annotated database on which networks can be trained. This assumption is not always satisfactory, notably when the domain where the network is transferred dramatically differs from the original one.

We refer the reader to [10] for a detailed overview concerning learning methods tailored explicitly for learning from scratch on small datasets. Some methods benefit from employing *geometric priors*, such as fixed or learnable filters based on wavelet transformations [47, 48, 19] or discrete cosine transform [59, 58]. Invariance to input transformations (e.g., rotation, translation) is achieved by integrating steerable filters or circular harmonics [65], alternative padding strategies [31], and specialized convolution blocks [66, 56]. Cost-based regularization strategies formulate objective functions and penalties to mitigate overfitting [45], such as the cosine loss and variants [4, 34, 56]. Other cost-based regularizers include rotation invariance [66], gradient penalties and spectral norms [7], low-rank embedding [38], and temperature calibration [8]. Another set of approaches performs data augmentation on the input space by relying on generative models [73, 72, 2, 51], or on the network's feature space [29, 32, 41, 42]. Finally, some previous work warm-start the final classifier after solving a pretext task through layer-wise greedy initialization [54], adap-

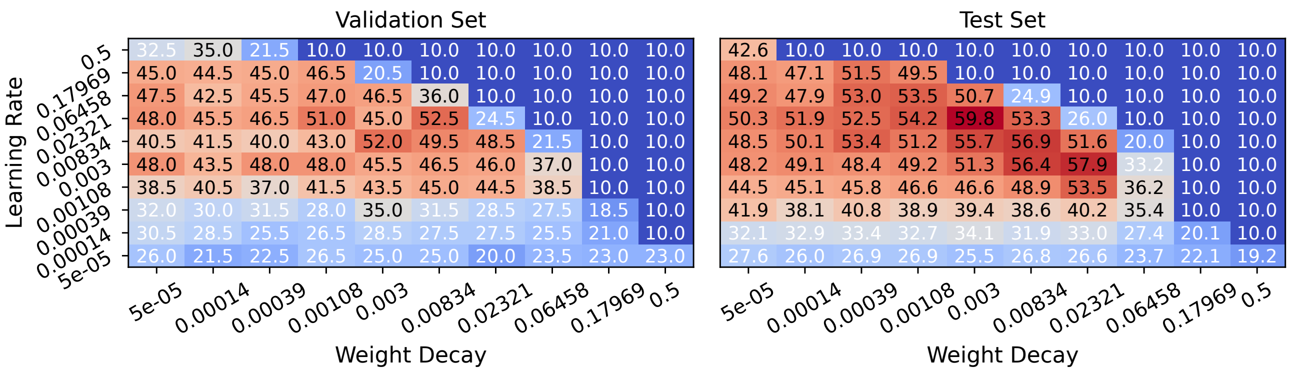

Figure 2: **Impact of small validation sets.** Validation and test accuracy scored by a WRN-16-1 trained with momentum. Having only available a small training set can result in sub-optimal model selection on noisy validation sets. In this case, the best model on the validation set does not transfer to the best model on the test set.

tive model complexity [18], dictionary-based learning [33], or self-supervised pre-training [74, 61].

Our work shares with previous work [9, 3] the interest in improving vanilla cross-entropy classifiers on limited data settings. Differently, we perform a comprehensive analysis concerning the impact of model size and training schedule length, which is completely missing in [9]. Further, we propose additional insights regarding the search for optimal optimization parameters and the impact of momentum.

3. Preliminaries

We face an image classification problem in which we are given a small set of N labeled pairs $\mathcal{D} = \{x_i, y_i\}_{i=1:N}$ sampled from distributions $\mathcal{X}$ and $\mathcal{Y}$. We train function approximators f_θ (WRNs) with mini-batches of dimension B to optimize the objective function $J_\theta = \frac{1}{B}\sum_{x,y\sim\mathcal{D}} J(f_\theta(x), y))$. The networks are trained for T iterations with SGD and its variants with momentum (μ) and weight decay (λ). The latter explicitly penalizes the L_2 squared norm of the weights divided by two. At each training step t, the parameters follow the update rule:

$$\begin{aligned} v_{t+1} &= \mu v_t + \alpha_t(\nabla J_\theta + \lambda\theta_t) \\ \theta_{t+1} &= \theta_t - v_{t+1} \end{aligned} \tag{1}$$

with α_t being the learning rate adjusted at each iteration step according to a defined learning rate schedule. We instead refer to α as the initial learning rate. If we consider the simpler case without momentum (i.e., $\mu = 0.0$), the general SGD update reported in Eq. (1) can be decoupled into a weight decay step $\theta_{t+1} = \theta_t(1-\alpha_t\lambda)$ and a gradient descent one $\theta_{t+1} = \theta_t - \alpha_t\nabla J_\theta$. The weight decay update is ruled by the *effective weight decay*, which is the product between α_t and λ [21]. If we assume scale-invariance[1], i.e., $J_\theta = J_{c\cdot\theta}, c > 0$, it follows that $\nabla J_\theta \cdot \theta = 0$ [60, 40]. Hence, each SGD step encompasses a combination of two conflicting forces. The *effective weight decay* diminishes the parameter norm, whereas the gradient amplifies it, resulting in a dynamic interplay between the two.

4. Experiments

To perform our empirical analyses, we choose the popular WRN architecture of depth 16 widely used in previous work on the small ciFAIR-10 dataset [47, 9], and vary the width to increase model size when necessary. We fix the batch size B for all the training runs to 10, given the success of small batches in small-data regimes [9, 10]. In addition, we incorporate the widely used cosine annealing schedule to adjust the learning rate during training [43]. To have a good glimpse of the impact of the learning rate and weight decay on the generalization performance, for most of the networks, we run grid searches with 100 models, sampling equally spaced learning rate and weight decay values in log-space from the interval $[5\cdot10^{-5}, 5\cdot10^{-1}]$. We only run a sub-portion of the grid for bigger models that would have required an onerous amount of compute. We finally employ minimal data augmentation composed of random horizontal flipping and translations of 4 pixels.

4.1. Baseline setup

As a base setup, we choose i) the smallest architecture of the WRN-16 family, i.e., WRN-16-1; ii) a training schedule of $25k$ steps as proposed in [9]; iii) momentum $\mu = 0.9$ as standard practice in deep learning; iv) model selection on a small validation set with the aforementioned grid search. In particular, we employ the training-validation split proposed in [9].

[1] All layers of WRNs are scale-invariant except for the BN affine parameters and final classification head.

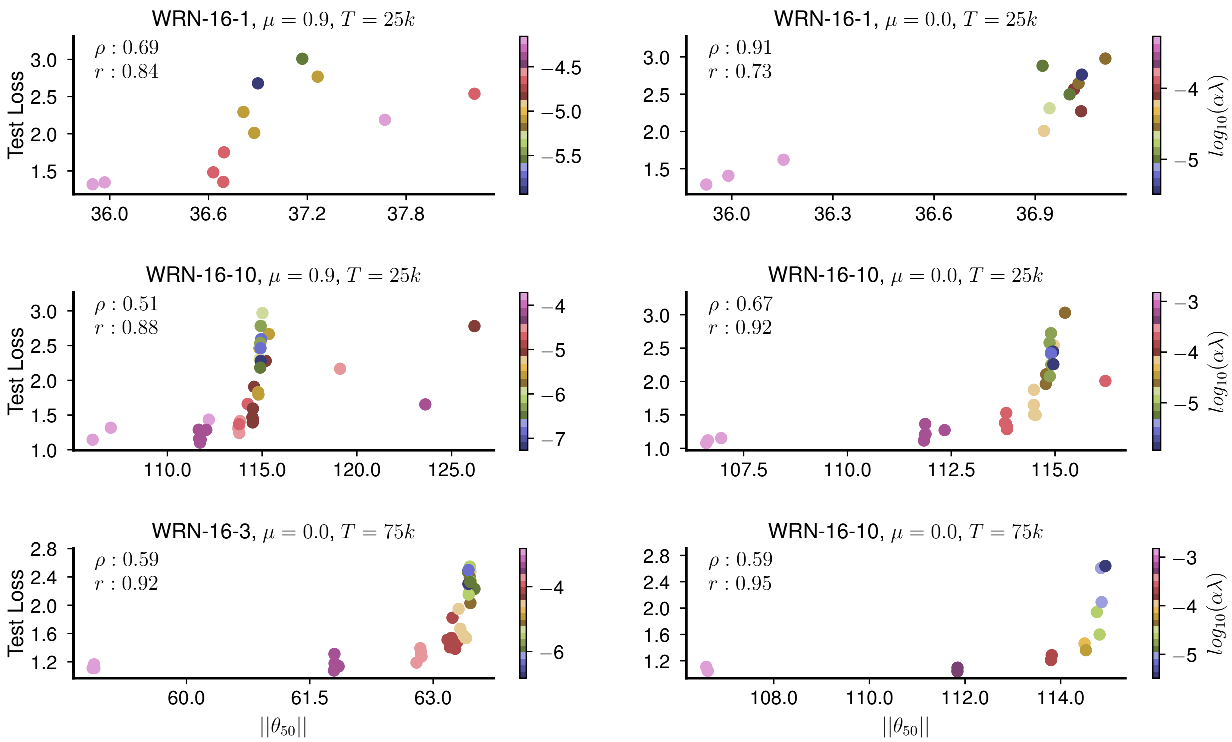

Figure 3: **Model selection via parameters norm.** Relationship between the norm of the parameters after one training epoch (50 iterations) and test loss. All networks have scored 100% training accuracy on the training set, μ represents momentum, and T the total number of training steps. Pearson and Spearman's coefficients are indicated with p and r.

In Fig. 2, we show the results of the grid searches for both validation and test sets. Given the accuracy score on the validation set, we select the model scoring around 53.3% on the testing set. However, we note two interesting insights: 1) the best learning rate and weight decay combination found in the validation set does not transfer to the optimal model, and 2) the best-achieved accuracy on the test set is already higher than previously published results of larger networks, e.g., WRN-16-8 [47, 58, 9, 10]. Given the limited size of the validation set, it is reasonable to believe that the model selection process may be noisy and suboptimal. Hence, a more targeted model selection could deliver networks that generalize better, particularly for larger models, as we have just observed that a tiny WRN-16-1 coupled with optimal parameters could outperform the best accuracy of a larger WRN-16-8.

4.2. HPs selection without validation sets

We devise a straightforward heuristic that effectively predicts the generalization performance of models by only monitoring training-related metrics. In this manner, we circumvent the requirement of relying on held-out validation sets, which may be limited and noisy in small-sample regimes. We first filter out all networks that do not fit the training set, i.e., those that do not score 100%. We are not interested in those networks since their low performance on the training set directly translates to poor generalization. Secondly, out of this pool of models, we consider the parameter vector norm $||\theta_t||$ at the beginning of training to be a good predictor for the testing loss. Previous work supports our intuitive approach by showing that regularization (e.g., weight decay) mostly affects early training dynamics [20].

In Fig. 3, we plot the test loss as a function of the norm after one epoch, which coincides with as few as 50 steps, i.e., $||\theta_{50}||$. We represent models that share the same learning rate-weight decay product in the same colors. A robust monotonic relationship exists between the two variables, as indicated by Spearman's rank coefficient surpassing 0.8 most of the time. The models with the smallest norm are the ones that generalize better by scoring lower testing losses. The monotonicity increases as the model size and training length increase. Reasonably, models with similar initial *effective weight decay* share norm magnitudes since their parameter vector is equally decayed. The symmetries across the learning rate-weight decay space (left-to-right diagonals) are also visible in Fig. 2 (right). However, not all the models generalize the same along a constant $\alpha\lambda$ since the gradient update is proportional to only α, not $\alpha\lambda$. Mo-

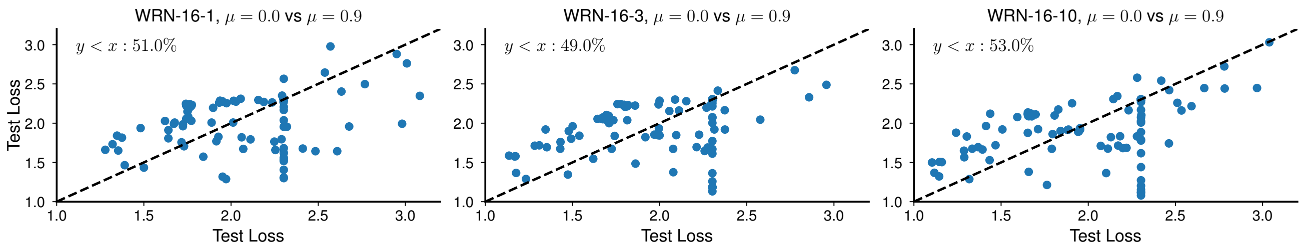

Figure 4: **Impact of momentum.** Comparison among three couples of architectures in terms of testing loss without ($\mu = 0.0$) and with ($\mu = 0.9$) momentum. The losses of networks trained without momentum are shown on the y-axis. Momentum does not seem to provide clear benefits in the optimization leading to similarly performing networks.

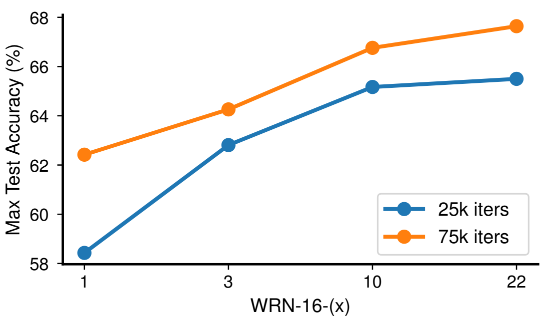

Figure 5: **Impact of training length.** Maximal achievable test accuracy as a function of the employed architecture and number of training iterations. A longer training schedule improves generalization.

mentum introduces some additional noise, potentially attributable to the more complex dynamics of incorporating previous gradients. However, the monotonic relationship remains reliable also if $\mu = 0.9$.

By using the parameter's norm to select the HPs, we raise the accuracy of the base WRN-16-1 from 53.3% to 56.9%. We will use this model-selection strategy in the next experiments.

4.3. Removal of momentum

Momentum is widely used in the deep learning community. Recent work has shown that it reduces the distance traveled by the parameters over the loss landscape [24] Furthermore, momentum makes the training dynamics slightly more complex due to past-gradients additions. We conducted experiments to assess the effect of momentum in our constrained data conditions using three models: WRN-16 with width scales of 1, 3, and 10. All six models underwent training for 25,000 steps. The test losses for each architecture, both with and without momentum, were compared, as depicted in Fig. 4. Remarkably, approximately 50% of the time, the best models are either with $\mu = 0.0$ or $\mu = 0.9$, indicating a similar test performance. These results suggest that making the SGD trajectories noisier may not necessarily penalize learning in limited data scenarios. To this end, we remove momentum and maintain more predictable training dynamics. In this manner, our momentum-free WRN-16-1 reaches a test accuracy of 58.1%, higher than the previous 56.9%. Notably, by removing momentum and performing HPs selection with our newly introduced metric (parameters' norm), we made a small WRN-16-1 as data-efficient as a larger WRN-16-8 tuned with Asynchronous HyperBand with Successive Halving (ASHA) search, which scored on the same benchmark 58.2% test accuracy [9].

4.4. Increased model size

Scaling up model size is a popular way to improve generalization [23]. However, with limited data, scaling the model without providing the right amount of regularization easily leads to overfitting. To better analyze the impact of scale, we report the test accuracy of WRN-16-1, WRN-16-3, and WRN-16-10, all trained without momentum in Fig. 6.

Increasing the width by $3\times$ already provides a maximum increase of 4.4 percentage points. The best achievable accuracy rises from 62.8% with WRN-16-3 to 65.2% with WRN-16-10. Our HPs selection metric correctly predicts the optimal learning rate-weight decay combination, and hence we gain 7.1 percent points to reach 65.2% test accuracy from the previous 58.2%.

4.5. Increased training length

Prior empirical evidence indicates that extended training schedules have demonstrated comparable performance to pre-trained networks. [22]. The limited data in small-sample scenarios bears the risk of under-training networks if the number of epochs and batch size are directly imported from the default setups with more data. Indeed, previous work showed that the number of training updates plays the most important role in learning [27].

To this end, we test a longer training schedule that

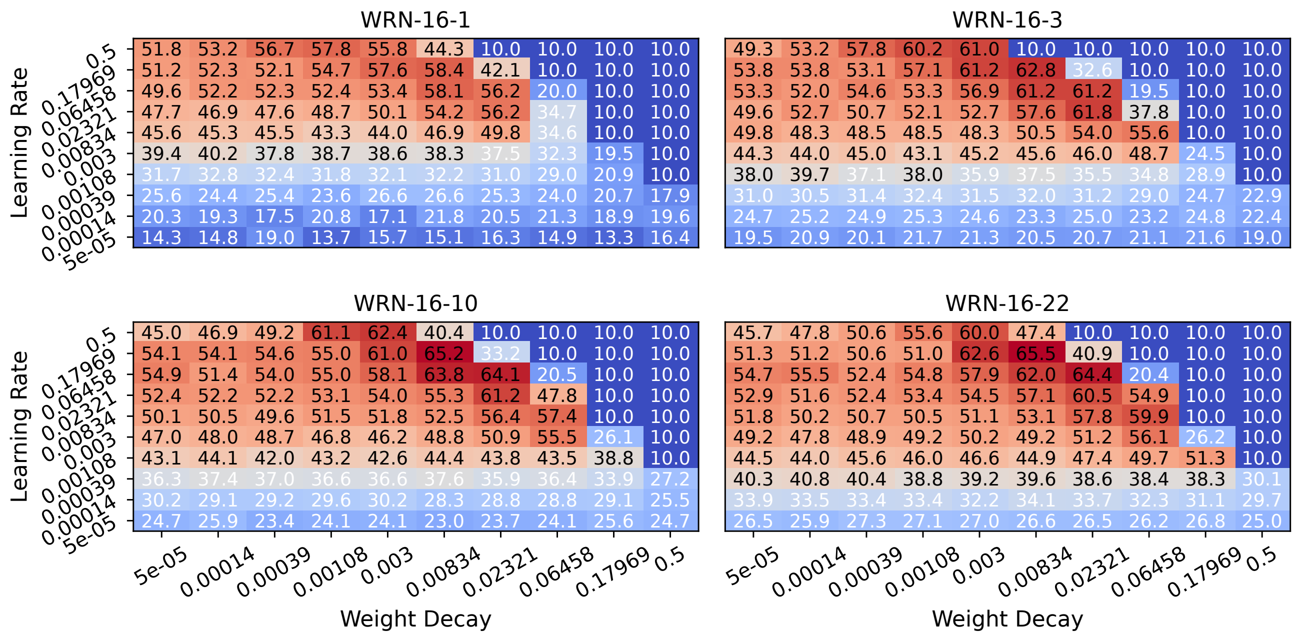

Figure 6: **Impact of model scale.** Test accuracy over the predefined learning rate-weight decay space. Increased model width significantly improves maximal achievable test accuracy but plateaus when moving from WRN-16-10 to WRN-16-22. All networks are trained for $25k$ iterations.

closely matches the one originally proposed in the paper that introduced the WRN architecture [69]. In particular, WRNs were trained on 50,000 samples for 200 epochs and mini-batches of size 128, resulting in a training schedule of $\sim 78k$ steps. To match this length, we triplicate the number of epochs from 500 to 1,500 while maintaining the batch size of dimension 10 to get a total of $75k$ training steps.

At all model scales, the tested networks improve their testing accuracy. In particular, the smallest WRN obtained the highest gain of 4 percent points. Not negligible improvements of 1.5, 1.6, and 2.1 percent points are scored by networks of widths 3,10 and 22, respectively.

The findings depicted in Fig. 5 reveal that the previously employed schedules are inadequately short. We also tested a longer training schedule of 4,500 epochs for the WRN-16-1 in preliminary experiments. We have not obtained significant improvements and hence stopped at 3,000. However, we do not rule out that increased training time could provide additional but moderate gains at large model scales.

Our final architecture becomes the WRN-16-22 trained for $75k$ iterations. The model selection strategy predicts the second-best model, which slightly underperforms the highest-scoring one (66.5% vs 67.6%). Increasing the model width from 10 to 22 and tripling the training length make us gain 1.3 percent points over the previous setup.

Pub.	Architecture	Augmentation	Accuracy
[9]	WRN-16-8	plain	58.22
[19]	Scatt. WRN	AutoAugment	$63.13 \pm 0.29^{*}$
[51]	WRN-16-8	MixUp	66.16 ± 0.78
[51]	WRN-16-8	ChimeraMix[1]	65.83 ± 0.78
[51]	WRN-16-8	ChimeraMix[2]	67.30 ± 1.21
Ours	WRN-16-10	plain	65.9
Ours	WRN-16-22	plain	66.5

Table 1: **Comparison with state-of-the-art methods.** All networks are trained on CIFAR-10 with 50 samples per class. *Scattering WRN has 22.6M parameters and is evaluated on the CIFAR-10 test set rather than ciFAIR-10. ChimeraMix[1] employs a grid-based patch selection while ChimeraMix[2] a gradient-based methodology. Plain augmentation is composed of simple horizontal flipping and 4-pixel translations.

4.6. Comparison with the state of the art

In the preceding sections, we tested and discussed several design choices to enhance our training scheme's overall performance without relying on hand-crafted data augmentations or costly generative models.

To gauge the effectiveness of our approach, we now compare our WRN-16-22 against the best state-of-the-art meth-

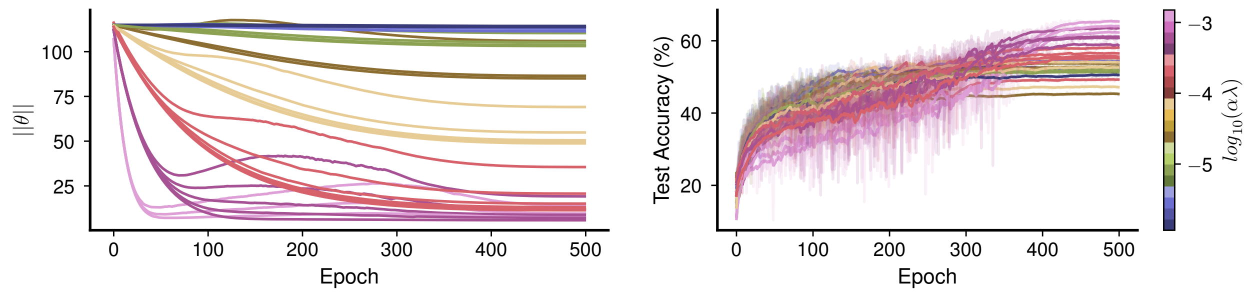

Figure 7: **Norm and test accuracy evolution.** We show the evolution of the weights norm and test accuracy for the WRN-16-10, which reached 100% training accuracy and was trained for 500 epochs ($25k$ iterations). The models with the highest $\alpha\lambda$ products experience more chaotic training dynamics (noisy test accuracy profile), fast decay of parameters norm, and better generalization.

ods. In particular, we benchmark against WRN-16-8 architectures trained with cross-entropy loss [9] plus basic, i.e., translation and horizontal flipping, or strong data augmentation methods such as MixUp [71] or ChimeraMix [51]. The hyper-parameters, i.e., learning rate and weight decay, were selected through ASHA search in the above cases. We also report the performance of recent parametric scattering networks [19] powered with AutoAugment [15].

We show the results in Table 1. Our WRN-16-10 and WRN-16-22 architectures trained with our scheme achieve recognition performance on par with ChimeraMix and MixUp and significantly outperform the WRN-16-8 from [9] and scattering networks coupled with AutoAugment [19]. Our reliance on plain data augmentation and implicit regularization techniques proves advantageous, as it enables our solution to generalize effectively across various domains, enhancing its practicality and transferability. Furthermore, our scheme could be theoretically coupled with such powerful data augmentation techniques if the image domain is agnostic to the biases introduced by handcrafted augmentations or if enough computational resources are available to train generative models properly.

4.7. Additional Analyses

Importance of HPs selection. We highlight that properly selecting hyper-parameters, particularly weight decay, is fundamental to providing optimal performance. For instance, referring to Fig. 6, if the value of weight decay is set too small ($5 \cdot 10^{-5}$), and a line search is performed over the learning rate, the maximum test accuracy improvement among WRN-16-1 and WRN-16-22 is approximately three percentage points. On the other hand, if the search is also expanded over the weight decay direction, the gain almost doubles to 7 percentage points.

Chaotic train dynamics generalize better. In Fig. 7, we provide additional insights regarding the evolution of the parameters norm and generalization through the test accuracy in the case of WRN-16-10 trained for $25k$ iterations. The largest weight decay-learning rate combinations that manage to fit the training set cause a fast decay of the parameters norm (as studied in Section 4.1) and also chaotic training dynamics. The right plot of Fig. 7 shows that a high $\alpha\lambda$ combination generates noisy test accuracy profiles and late convergence. Our findings align with previous studies [39, 30, 35], which suggest that training with higher learning rates leads to solutions with improved sharpening and generalization profiles.

HPs transfer across model sizes. Interestingly, it is also visible that the difference in parameter norm at the start of training due to increased model size drifts the area of better generalization towards the bottom right. This is partially explainable because the weight decay, as mentioned in Section 4.1, directly scales the weight vector by $\alpha_t\lambda$ while the gradient update does not depend on the parameter norm but just the learning rate. Consequently, when the weight norm increases, the gradient step becomes smaller than the weight decay update. However, as visible in Fig. 6, the best HPs combination remains constant across sizes, although the number of parameters has increased from 0.17M of WRN-16-1 to approximately 82.73M of WRN-16-22. Further investigations are necessary to gain a deeper understanding of this phenomenon. The consistency of optimal HPs presents a promising avenue for future research, offering potential computational savings and improved efficiency.

5. Conclusions

In this work, we presented and ablated a simple methodology to push the limits of classifier recognition performance with small training datasets in image classification.

While approaches based on aggressive data augmentation and generative models can raise classification abilities through data synthesis, they still have limitations, such as being domain-specific or requiring extensive computational resources and careful design. On the other hand, we explored several factors to improve the model's performance with alternative regularizations, including selecting optimal HPs more reliably and scaling the model size and training schedule. By implementing these techniques, we achieved state-of-the-art performance on the popular ciFAIR-10 small-data benchmark, demonstrating the validity of our empirical analyses.

Although tested on a single dataset, our work provides valuable insights that can benefit practitioners and researchers interested in developing strategies to improve generalization in small-data settings.

References

[1] Madhu S Advani, Andrew M Saxe, and Haim Sompolinsky. High-dimensional dynamics of generalization error in neural networks. *Neural Networks*, 2020. 2

[2] Idan Azuri and Daphna Weinshall. Generative latent implicit conditional optimization when learning from small sample. In *2020 25th International Conference on Pattern Recognition (ICPR)*, 2021. 1, 2

[3] Björn Barz, Lorenzo Brigato, Luca Iocchi, and Joachim Denzler. A strong baseline for the vipriors data-efficient image classification challenge. *arXiv preprint arXiv:2109.13561*, 2021. 3

[4] Björn Barz and Joachim Denzler. Deep learning on small datasets without pre-training using cosine loss. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2020. 2

[5] Björn Barz and Joachim Denzler. Do we train on test data? purging CIFAR of near-duplicates. *Journal of Imaging*, 6(6), 2020. 2

[6] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias-variance trade-off. *Proceedings of the National Academy of Sciences*, 2019. 2

[7] Alberto Bietti, Grégoire Mialon, Dexiong Chen, and Julien Mairal. A kernel perspective for regularizing deep neural networks. In *International Conference on Machine Learning (ICML)*, 2019. 1, 2

[8] Jorg Bornschein, Francesco Visin, and Simon Osindero. Small data, big decisions: Model selection in the small-data regime. In *International Conference on Machine Learning*, 2020. 2

[9] Lorenzo Brigato, Björn Barz, Luca Iocchi, and Joachim Denzler. Tune it or don't use it: Benchmarking data-efficient image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1071–1080, 2021. 1, 2, 3, 4, 5, 6, 7

[10] Lorenzo Brigato, Björn Barz, Luca Iocchi, and Joachim Denzler. Image classification with small datasets: overview and benchmark. *IEEE Access*, 2022. 1, 2, 3, 4

[11] Lorenzo Brigato and Luca Iocchi. A close look at deep learning with small data. In *2020 25th International Conference on Pattern Recognition (ICPR)*, 2021. 2

[12] Robert-Jan Bruintjes, Attila Lengyel, Marcos Baptista Rios, Osman Semih Kayhan, and Jan van Gemert. VIPriors 1: Visual inductive priors for data-efficient deep learning challenges. *arXiv preprint arXiv:2103.03768*, 2021. 1

[13] Robert-Jan Bruintjes, Attila Lengyel, Marcos Baptista Rios, Osman Semih Kayhan, Davide Zambrano, Nergis Tomen, and Jan van Gemert. Vipriors 3: Visual inductive priors for data-efficient deep learning challenges. *arXiv preprint arXiv:2305.19688*, 2023. 1

[14] Alon Brutzkus and Amir Globerson. Why do larger models generalize better? a theoretical perspective via the xor problem. In *International Conference on Machine Learning*, 2019. 2

[15] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501*, 2018. 1, 7

[16] Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International conference on machine learning*, 2019. 2

[17] Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. *arXiv preprint arXiv:1810.02054*, 2018. 2

[18] Jiashi Feng and Trevor Darrell. Learning the structure of deep convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, 2015. 3

[19] Shanel Gauthier, Benjamin Thérien, Laurent Alsene-Racicot, Muawiz Chaudhary, Irina Rish, Eugene Belilovsky, Michael Eickenberg, and Guy Wolf. Parametric scattering networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2, 6, 7

[20] Aditya Sharad Golatkar, Alessandro Achille, and Stefano Soatto. Time matters in regularizing deep networks: Weight decay and data augmentation affect early learning dynamics, matter little near convergence. *Advances in Neural Information Processing Systems*, 2019. 4

[21] Stephen Hanson and Lorien Pratt. Comparing biases for minimal network construction with back-propagation. *Advances in neural information processing systems*, 1988. 3

[22] Kaiming He, Ross Girshick, and Piotr Dollár. Rethinking imagenet pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019. 2, 5

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2, 5

[24] Byeongho Heo, Sanghyuk Chun, Seong Joon Oh, Dongyoon Han, Sangdoo Yun, Gyuwan Kim, Youngjung Uh, and Jung-Woo Ha. Adamp: Slowing down the slowdown for momentum optimizers on scale-invariant weights. *arXiv preprint arXiv:2006.08217*, 2020. 5

[25] Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Mostofa Ali

Patwary, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017. 2

[26] Elad Hoffer, Ron Banner, Itay Golan, and Daniel Soudry. Norm matters: efficient and accurate normalization schemes in deep networks. *Advances in Neural Information Processing Systems*, 2018. 2

[27] Elad Hoffer, Itay Hubara, and Daniel Soudry. Train longer, generalize better: closing the generalization gap in large batch training of neural networks. *Advances in neural information processing systems*, 2017. 2, 5

[28] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, 2015. 2

[29] Masato Ishii and Atsushi Sato. Training deep neural networks with adversarially augmented features for small-scale training datasets. In *2019 International Joint Conference on Neural Networks (IJCNN)*, 2019. 2

[30] Nikhil Iyer, V Thejas, Nipun Kwatra, Ramachandran Ramjee, and Muthian Sivathanu. Wide-minima density hypothesis and the explore-exploit learning rate schedule. *Journal of Machine Learning Research*, 2023. 7

[31] Osman Semih Kayhan and Jan C van Gemert. On translation invariance in CNNs: Convolutional layers can exploit absolute spatial location. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 2

[32] Rohit Keshari, Richa Singh, and Mayank Vatsa. Guided dropout. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019. 2

[33] Rohit Keshari, Mayank Vatsa, Richa Singh, and Afzel Noore. Learning structure and strength of cnn filters for small sample size training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 3

[34] Takumi Kobayashi. T-vMF similarity for regularizing intraclass feature distribution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[35] Maxim Kodryan, Ekaterina Lobacheva, Maksim Nakhodnov, and Dmitry P Vetrov. Training scale-invariant neural networks on the sphere can happen in three regimes. *Advances in Neural Information Processing Systems*, 2022. 2, 7

[36] Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better imagenet models transfer better? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2661–2671, 2019. 2

[37] Attila Lengyel, Robert-Jan Bruintjes, Marcos Baptista Rios, Osman Semih Kayhan, Davide Zambrano, Nergis Tomen, and Jan van Gemert. Vipriors 2: visual inductive priors for data-efficient deep learning challenges. *arXiv preprint arXiv:2201.08625*, 2022. 1

[38] José Lezama, Qiang Qiu, Pablo Musé, and Guillermo Sapiro. OLÉ: Orthogonal low-rank embedding-a plug and play geometric loss for deep learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[39] Yuanzhi Li, Colin Wei, and Tengyu Ma. Towards explaining the regularization effect of initial large learning rate in training neural networks. *Advances in Neural Information Processing Systems*, 2019. 7

[40] Zhiyuan Li, Kaifeng Lyu, and Sanjeev Arora. Reconciling modern deep learning with traditional optimization analyses: The intrinsic learning rate. *Advances in Neural Information Processing Systems*. 2, 3

[41] Luyue Lin, Bo Liu, Xin Zheng, and Yanshan Xiao. An efficient image categorization method with insufficient training samples. *IEEE Transactions on Cybernetics*, 2020. 2

[42] Luyue Lin, Dacai Liu, Bo Liu, and Yanshan Xiao. A latent variables augmentation method based on adversarial training for image categorization with insufficient training samples. In *2020 16th International Conference on Control, Automation, Robotics and Vision (ICARCV)*, 2020. 2

[43] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations (ICLR)*, 2017. 3

[44] Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment*, 2021. 2

[45] Aviv Navon, Idan Achituve, Haggai Maron, Gal Chechik, and Ethan Fetaya. Auxiliary learning by implicit differentiation. In *International Conference on Learning Representations*, 2021. 1, 2

[46] Roman Novak, Yasaman Bahri, Daniel A Abolafia, Jeffrey Pennington, and Jascha Sohl-Dickstein. Sensitivity and generalization in neural networks: an empirical study. *arXiv preprint arXiv:1802.08760*, 2018. 2

[47] Edouard Oyallon, Eugene Belilovsky, and Sergey Zagoruyko. Scaling the scattering transform: Deep hybrid networks. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. 1, 2, 3, 4

[48] Edouard Oyallon, Sergey Zagoruyko, Gabriel Huang, Nikos Komodakis, Simon Lacoste-Julien, Matthew Blaschko, and Eugene Belilovsky. Scattering networks for hybrid representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2018. 2

[49] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 2009. 2

[50] Mohammad Pezeshki, Amartya Mitra, Yoshua Bengio, and Guillaume Lajoie. Multi-scale feature learning dynamics: Insights for double descent. In *International Conference on Machine Learning*, 2022. 2

[51] Christoph Reinders, Frederik Schubert, and Bodo Rosenhahn. Chimeramix: Image classification on small datasets via masked feature mixing. In Luc De Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022*, 2022. 1, 2, 6, 7

[52] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*, 2021. 1

[53] Jonathan S Rosenfeld, Amir Rosenfeld, Yonatan Belinkov, and Nir Shavit. A constructive prediction of the generalization error across scales. *arXiv preprint arXiv:1909.12673*, 2019. 2

[54] Diego Rueda-Plata, Raúl Ramos-Pollán, and Fabio A González. Supervised greedy layer-wise training for deep convolutional networks with small datasets. In *Computational Collective Intelligence*. 2015. 2

[55] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 1

[56] Pengfei Sun, Xuan Jin, Wei Su, Yuan He, Hui Xue, and Quan Lu. A visual inductive priors framework for data-efficient image classification. In *European Conference on Computer Vision (ECCV) Workshops*, pages 511–520, Cham, 2020. Springer International Publishing. 2

[57] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, 2019. 2

[58] Matej Ulicny, Vladimir A Krylov, and Rozenn Dahyot. Harmonic networks for image classification. In *BMVC*, 2019. 2, 4

[59] Matej Ulicny, Vladimir A Krylov, and Rozenn Dahyot. Harmonic networks with limited training samples. In *European Signal Processing Conference (EUSIPCO)*, 2019. 2

[60] Twan Van Laarhoven. L2 regularization versus batch and weight normalization. *arXiv preprint arXiv:1706.05350*, 2017. 2, 3

[61] Tan Wad, Qianru Sun, Sugiri Pranata, Karlekar Jayashree, and Hanwang Zhang. Equivariance and invariance inductive bias for learning from insufficient data. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XI*, 2022. 3

[62] Ruosi Wan, Zhanxing Zhu, Xiangyu Zhang, and Jian Sun. Spherical motion dynamics: Learning dynamics of normalized neural network using sgd and weight decay. *Advances in Neural Information Processing Systems*, 2021. 2

[63] Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 2018. 2

[64] Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM Computing Surveys (CSUR)*, 2020. 2

[65] Daniel E Worrall, Stephan J Garbin, Daniyar Turmukhambetov, and Gabriel J Brostow. Harmonic networks: Deep translation and rotation equivariance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

[66] Wenju Xu, Guanghui Wang, Alan Sullivan, and Ziming Zhang. Towards learning affine-invariant representations via data-efficient cnns. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020. 2

[67] Juseung Yun, Byungjoo Kim, and Junmo Kim. Weight decay scheduling and knowledge distillation for active learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI 16*, 2020. 2

[68] Juseung Yun, Janghyeon Lee, Hyounguk Shon, Eojindl Yi, Seung Hwan Kim, and Junmo Kim. On the angular update and hyperparameter tuning of a scale-invariant network. In *European Conference on Computer Vision*, 2022. 2

[69] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In Edwin R. Hancock Richard C. Wilson and William A. P. Smith, editors, *British Machine Vision Conference (BMVC)*, pages 87.1–87.12. BMVA Press, September 2016. 2, 6

[70] Guodong Zhang, Chaoqi Wang, Bowen Xu, and Roger Grosse. Three mechanisms of weight decay regularization. *arXiv preprint arXiv:1810.12281*, 2018. 2

[71] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 7

[72] Xiaofeng Zhang, Zhangyang Wang, Dong Liu, Qifeng Lin, and Qing Ling. Deep adversarial data augmentation for extremely low data regimes. *IEEE Transactions on Circuits and Systems for Video Technology*, 2020. 2

[73] Xiaofeng Zhang, Zhangyang Wang, Dong Liu, and Qing Ling. Dada: Deep adversarial data augmentation for extremely low data regime classification. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019. 1, 2

[74] Bingchen Zhao and Xin Wen. Distilling visual priors from self-supervised learning. In *European Conference on Computer Vision*, 2020. 3