When Dynamic Data Selection Meets Data Augmentation: Achieving Enhanced Training Acceleration

Suorong Yang [1 2] **Peng Ye** [2 3] **Furao Shen*** [1] **Dongzhan Zhou*** [2]

Abstract

Dynamic data selection aims to accelerate training with lossless performance. However, reducing training data inherently limits data diversity, potentially hindering generalization. While data augmentation is widely used to enhance diversity, it is typically not optimized in conjunction with selection. As a result, directly combining these techniques fails to fully exploit their synergies. To tackle the challenge, we propose a novel online data training framework that, for the first time, unifies dynamic data selection and augmentation, achieving both training efficiency and enhanced performance. Our method estimates each sample's joint distribution of local density and multimodal semantic consistency, allowing for the targeted selection of augmentation-suitable samples while suppressing the inclusion of noisy or ambiguous data. This enables a more significant reduction in dataset size without sacrificing model generalization. Experimental results demonstrate that our method outperforms existing state-of-the-art approaches on various benchmark datasets and architectures, e.g., reducing 50% training costs on ImageNet-1k with lossless performance. Furthermore, our approach enhances noise resistance and improves model robustness, reinforcing its practical utility in real-world scenarios.

1. Introduction

Deep learning has thrived with the growing availability of large-scale datasets. As models become more complex and parameter-extensive, it is necessary to utilize even larger datasets, introducing challenges like reduced training efficiency. Moreover, large-scale datasets often include redundant or noisy samples (Xia et al., 2023b; Northcutt et al., 2021; Wang et al., 2018), which can compromise training effectiveness. To address these challenges, various data selection strategies have been proposed to reduce dataset size and enhance data efficiency while maintaining model performance. These methods can be broadly categorized into static data selection (Tan et al., 2024; Zhang et al., 2024; Xia et al., 2023b) and dynamic data selection (or pruning) (Qin et al., 2024; Raju et al., 2021; He et al., 2024). Static selection identifies a fixed subset of data before training begins, whereas dynamic pruning continuously selects the most influential samples during training. While these methods can effectively reduce training costs without degrading performance, the reduced training data volume often leads to reduced data diversity. Consequently, the generalization of deep models is limited, and lossless performance is typically achieved with relatively high selection ratios.

In fact, to address the issue of data diversity, data augmentation is commonly used in model training, which can also improve generalization (Xu et al., 2023; Yang et al., 2022b). However, effectively combining data selection and augmentation remains a challenge. Existing data selection methods typically prioritize representative and challenging samples, which are not specifically designed with augmentation in mind. Although augmentation can increase the diversity of selected data and further improve model robustness, applying it to complex samples may introduce ambiguity or noise, potentially increasing training difficulty (Gong et al., 2021; Yang et al., 2024b). Therefore, integrating data selection and augmentation in a unified framework presents a promising yet underexplored direction to balance efficiency and generalization.

In this study, we propose a novel framework that integrates dynamic data selection with augmentation, enabling a unified approach to enhance training efficiency and model generalization. During model training, low-density samples often correspond to underlearned or insufficiently represented data points, such as classification boundaries. Applying augmentation transformations to these samples reinforces model learning and improves robustness. However, noisy or outlier samples typically exhibit relatively low density, increasing the risk of introducing noise. To address this, we introduce a semantic consistency distribution derived from

[1]National Key Laboratory for Novel Software Technology, Nanjing University [2]Shanghai Artificial Intelligence Laboratory [3]The Chinese University of Hong Kong.

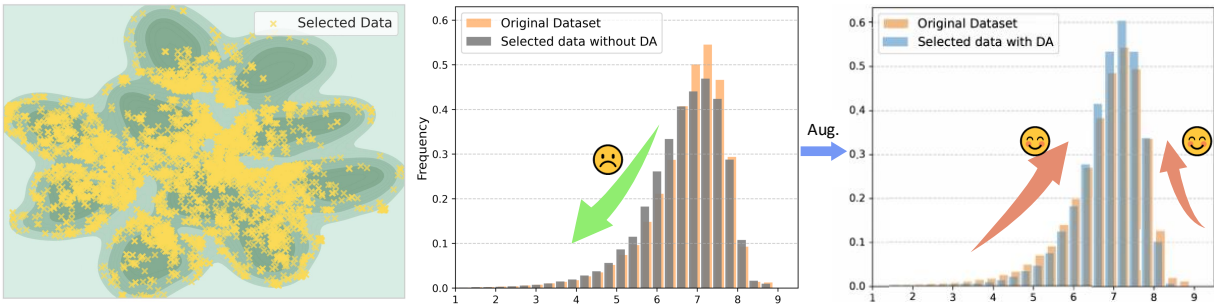


Figure 1. Illustration of the distribution of our selected data points using T-SNE algorithm (*left*) and the density histograms without (*mid*) and with (*right*) augmenting the selected data on CIFAR-10. The selection ratio is 10%.

the pre-trained multimodal model CLIP (Radford et al., 2021). By prioritizing samples with high sparsity and strong semantic alignment, our approach leverages the joint distribution of density and semantic consistency for effective and robust sample selection.

As illustrated in the left sub-figure of Fig. 1, the selected data points predominantly cluster around boundary regions among clusters. Meanwhile, the density histogram in Fig. 1 reveals a more balanced distribution after augmentation, with fewer low- and high-density samples and more data points converging toward the moderate-density regions. Compared to the distribution without DA, this redistribution highlights our framework's ability to enhance sparse regions, improving model generalization across the entire data distribution.

Experimental results across benchmark datasets and deep architectures demonstrate the effectiveness of our method. On large-scale datasets such as Tiny-ImageNet (Chrabaszcz et al., 2017) and ImageNet-1k (Deng et al., 2009), our method significantly accelerates training while maintaining or even improving generalization. For instance, on ImageNet-1k, our approach doubles the training efficiency while achieving comparable performance with the entire dataset. Moreover, our framework exhibits strong cross-architecture and cross-scenario generalization, effectively mitigating the impact of noisy data and enhancing versatility in real-world applications. On Tiny-ImageNet, our approach outperforms leading baselines by at least 3% in accuracy under noisy conditions, further demonstrating its reliability.

Our main contributions are summarized as follows: 1) We propose a novel training framework that dynamically integrates data selection and augmentation, significantly accelerating training while maintaining model performance. 2) We introduce a joint distribution based on density and semantic consistency, ensuring effective sample selection and reducing noise and ambiguity. 3) Extensive experiments across diverse datasets and architectures demonstrate superior accuracy and generalization ability, particularly in noisy and challenging scenarios, validating its practical applicability.

2. Related Work

2.1. Data Selection

The primary goal of data selection is to enhance data-efficient learning, which can be broadly categorized into dataset distillation (Lei & Tao, 2023; Du et al., 2023; Sun et al., 2024; Liu & Wang, 2024), and static or dynamic data selection (or pruning) (Tan et al., 2024; Xia et al., 2023b; Sorscher et al., 2022; Qin et al., 2024). Dataset distillation focuses on synthesizing a small representative dataset that preserves the performance of training on the full dataset. In contrast, following dynamic data selection without synthesizing new data in this work, we propose a new data training framework that unifies dynamic data selection and augmentation to achieve enhanced model training acceleration.

Static data selection identifies a fixed subset of the training dataset before training begins. Existing methods can be categorized into selection with importance criteria, dataset distribution-based methods, and optimization-based methods. *Selection with importance criteria* computes per-sample importance scores and selects the most informative samples. This includes: 1) the expectation of ℓ_2-norm error vector and the gradient norm (EL2N and GraNd) (Paul et al., 2021), 2) the change in the optimal empirical risk when a sample is removed (Tan et al., 2024), 3) the number of forgetting events in the whole training process (Forgetting) (Toneva et al., 2018), 4) the impact of including or excluding a sample on the model's classification ability (Feldman & Zhang, 2020). *Dataset distribution-based methods* select samples based on the geometric distribution of the dataset. Herding (Welling, 2009) chooses samples based on their distance from the corresponding class centers. D2 (Maharana et al., 2023) defines sample difficulty by incorporating the difficulty of its neighboring examples. The work (Ramalingam et al., 2023) applies greedy k-center to select the coreset with good data coverage, and CCS (Zheng et al., 2023) balances the sample distribution

and importance in selection. Similarly, Moderate-DS (Xia et al., 2023b) selects samples that are closer to the median score, aiming to balance diversity and representativeness. *Optimization-based methods* formulate selection as an optimization problem using techniques such as scalable self-supervised pruning metrics (Sorscher et al., 2022), influence function (Yang et al., 2023a), bi-level optimization (Killamsetty et al., 2021), gradient matching (Mirzasoleiman et al., 2020b), convex optimization (Mirzasoleiman et al., 2020a), facility location function (Yang et al., 2023b), temporal dual-depth scoring (Zhang et al., 2024), and submodularity (Iyer et al., 2021; Nohyun et al., 2023).

Dynamic data selection identifies informative samples throughout training, allowing the dataset to adapt as the model learns. The work (Raju et al., 2021) proposes UCB and ϵ-greedy algorithms to estimate the uncertainty value associated with each training sample, selecting a subset of the data that exhibits the highest levels of uncertainty. Similarly, the work (He et al., 2024) also employs both prediction uncertainty and training dynamics to guide the selection process, ensuring that the most informative samples are retained throughout training. The work (Liu & Mirzasoleiman, 2022) proposes a data-efficient framework for training neural networks and achieves promising results. SAS (Joshi & Mirzasoleiman, 2023) improves data efficiency in SSL by proving and selecting the most beneficial data for contrastive training. Moreover, InfoBatch (Qin et al., 2024) proposes a method for unbiased dynamic data selection that accelerates training by pruning less informative samples to retain their relevance in model optimization, which allows for more efficient training without compromising the model's performance.

2.2. Data Augmentation

Data augmentation (DA) improves the generalization of deep neural networks by increasing the diversity of training samples (Yang et al., 2024a). Existing DA methods can be divided into image erasing/mixing-based and automatic augmentation methods (Xu et al., 2023). Image erasing and mixing-based augmentation erase some sub-regions in images or mix random information from two or more images for augmentation to create new samples, respectively. These methods include Cutout (DeVries & Taylor, 2017), RandomErasing (Zhong et al., 2020), HaS (Singh & Lee, 2017), AdvMask (Yang et al., 2022a), Mixup (Zhang et al., 2018), GuidedMixup (Kang & Kim, 2023), and GradSalMix (Hong et al., 2023), etc. In addition, based on pre-defined or optimized image transformation policies, automatic DA methods randomly apply one or multiple transformations to each image at each epoch, including AutoAugment (Cubuk et al., 2019), Fast-AutoAugment (Lim et al., 2019), RandAugment (Cubuk et al., 2020), TrivialAugment (Müller & Hutter, 2021), SelectAugment (Lin et al., 2023), EntAugment (Yang et al., 2024b), and MADAug (Hou et al., 2023), etc. Beyond these, generative data augmentation further enriches data by synthesizing new samples using generative models (Moreno-Barea et al., 2020). Recent studies also emphasize representation consistency (Atienza, 2022) and address distribution gaps between clean and augmented data (He et al., 2019), pointing to new challenges in effective DA design.

3. The Proposed Method

3.1. Overview of the Proposed Method

As shown in Fig. 2, we propose a novel data training framework that integrates dynamic data selection with augmentation to enhance both training efficiency and generalization. Our framework employs two complementary distributions: 1). a density distribution, dynamically estimated by a task-specific model (e.g., ResNet), which identifies underrepresented samples for augmentation and training, and (2) a semantic consistency distribution, computed using a frozen pre-trained multimodal model (CLIP), which quantifies the alignment between an image and its corresponding textual label. Low-density regions highlight underrepresented samples but may also include noisy or ambiguous instances. To address this issue, the semantic consistency distribution acts as a strong complement, filtering samples with weak semantic alignment. By combining both distributions, we construct a joint distribution that captures the relationship between sample informativeness and semantic correctness. Consequently, samples with higher joint distribution scores are prioritized to be selected and augmented for training.

3.2. How to Identify Samples for Augmentation

Given a dataset D that follows an underlying distribution $P(D)$, the optimization objective of our dynamic data selection module at time t is to select a subset $\hat{D}_t$, containing at most k samples. The goal is to minimize the expected loss over the distribution $P(D)$, with the following optimization: $\hat{D}_t = \underset{\hat{D}_t \subseteq D, \left|\hat{D}_t\right| \leq k}{\arg\min} \mathbb{E}_{z \sim P(D)} \left[L\left(z, \hat{\theta}_{\mathcal{A}(\hat{D}_t)}\right) \right]$, where z represents a test sample, L is the loss function, $\hat{\theta}_{\hat{D}_t}$ is the empirical risk minimizer on $\hat{D}_t$, and $\mathcal{A}$ represents the augmentation operations applied to the selected subset.

Our method prioritizes low-density samples because such regions in the feature space often correspond to underlearned or insufficiently represented data points. Focusing on these samples and applying augmentation operations helps the model capture distinctive features. By augmenting these sparse samples, we compensate for their underrepresentation, improving the model's ability to generalize across diverse data regions. To efficiently determine the density of data points during online training, we exploit an online approximate nearest neighbor search architecture

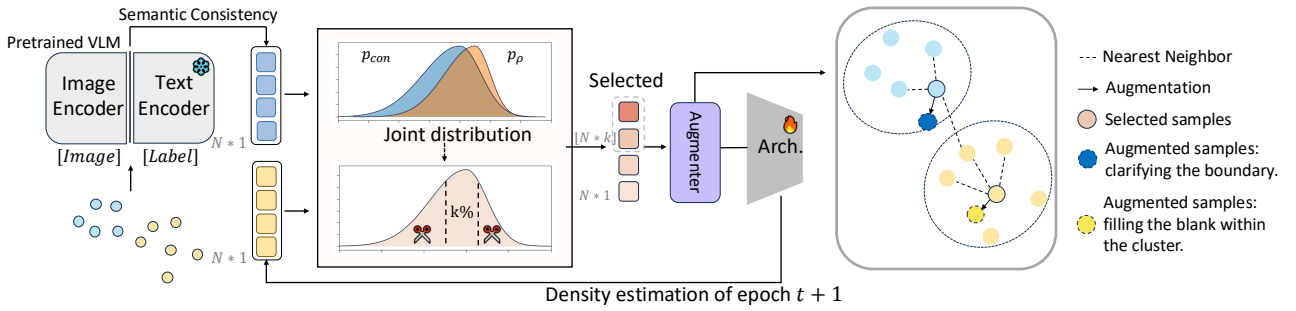


Figure 2. The framework of our proposed data training method: The core idea of our framework is to construct a joint distribution that integrates both the density and semantic consistency distributions, enabling the prioritization of low-density, semantically consistent samples. After augmentation, augmented sparse samples in intra-cluster regions help to fill the underrepresented spaces, while samples located around the decision boundaries between clusters differentiate the classification decision more clearly, thus improving generalization.

(HNSW) (Malkov & Yashunin, 2018) to query the nearest k neighbors of each sample x, denoted as $NN(x)$. The density of a sample is then estimated as the mean of the ℓ_2 distance between x and its neighbors:

$$\rho_{x_i} = \frac{1}{k} \sum_{j \in NN(x_i)} ||x_i - x_j||, \tag{1}$$

where higher values indicate low-density samples. The density scores are then normalized using a Min-Max scaling to obtain $p_\rho(x)$. For augmentation, we employ slight augmentation, which generates neighboring samples within low-density areas while preserving local structure.

Nevertheless, low-density regions may also contain challenging, outlier, or noisy samples. Continuously selecting these samples for training can substantially complicate training, especially on real-world datasets that inevitably include noise. To address this issue and improve the practical effectiveness of our method, we introduce a multimodal semantic consistency constraint that simultaneously refines the selection of low-density samples. This ensures that augmentation is applied to meaningful data, improving both robustness and efficiency in the training process.

3.3. Multimodal Consistency Estimation for Robust Data Selection

Noisy data - arising from incorrect labels, corrupted images, or outlier samples - reflects a fundamental mismatch between the semantic content of x and its corresponding label y. To detect and filter such inconsistencies, the data cleaner evaluates the joint distribution $p(x, y)$, which captures the plausibility of an image-label pair. Given the inherent correlation and multimodal nature of x and y, we introduce multimodal consistency supervision as an additional criterion for assessing sample reliability. This complements the density-based selector by filtering out samples that exhibit low cross-modal consistency.

To implement this, we leverage a pre-trained CLIP model to embed images and text into a shared multimodal space, enabling semantic alignment assessment. However, CLIP's zero-shot generalization is limited to domain-specific datasets, making it necessary to adapt the embeddings to the target domain. Instead of performing computationally expensive fine-tuning, we incorporate lightweight adapters of MLP for both the image and text encoders (Poth et al., 2023; Yang et al., 2024c). The lightweight architecture of adapters ensures efficient adaptation while preserving CLIP's pretrained knowledge.

To measure the cross-modal consistency, we compute the cosine similarity of the encoded image and text features:

$$con(x_i) = \ell_{cos}(E_I(x_i), E_T(y_i)), \tag{2}$$

where E_I and E_T are visual and textual encoders, respectively. The consistency scores are normalized via Min-Max scaling to approximate the consistency distribution $p_{con}(x)$, where higher values indicate stronger semantic alignment. Since the image-label alignment is derived from a pretrained vision-language model and remains independent of the training process, we precompute the consistency distribution beforehand. This enables direct use during sample selection, eliminating additional computational overhead in online training.

3.4. Augmenter

To integrate structural sparsity and semantic consistency, we define a joint distribution that combines both density and consistency distributions:

$$p_{sel}(x_i) = p_\rho(x_i) * p_{con}(x_i). \tag{3}$$

Here, p_ρ evolves dynamically with model training, allowing sample selection to adapt to the current model training state. During online training, samples with higher joint distribution scores are prioritized for augmentation and training,

Table 1. The accuracy (%) comparison to state-of-the-art baselines. All methods are trained with ResNet-18 on CIFAR-10/100 and ResNet-50 on Tiny-ImageNet. Note that some results could not be computed due to the unavailability of open-source code and parameter settings, making it impossible to reproduce. Random* means randomly selecting samples in each epoch.

Dataset	CIFAR-10			CIFAR-100			Tiny-ImageNet		
Whole Dataset	95.6			78.2			45.0		
Selection Ratio (%)	30	50	70	30	50	70	30	50	70
Random	90.2	92.3	93.9	69.7	72.1	73.8	29.8	37.2	42.2
Herding (Welling, 2009)	80.1	88.0	92.2	69.6	71.8	73.1	29.4	31.6	39.8
EL2N (Paul et al., 2021)	91.6	95.0	95.2	69.5	72.1	77.2	26.6	37.1	44.0
GraNd (Paul et al., 2021)	91.2	94.6	95.3	68.8	71.4	74.6	29.7	36.3	43.2
Glister (Killamsetty et al., 2021)	90.9	94.0	95.2	70.4	73.2	76.6	30.1	39.5	43.9
Forgetting (Toneva et al., 2018)	91.7	94.1	94.7	69.9	73.1	75.3	28.7	33.0	41.2
Moderate-DS (Xia et al., 2023b)	91.5	94.1	95.2	70.2	73.4	77.3	30.6	38.2	42.8
Self-sup. prototypes (Sorscher et al., 2022)	91.0	94.0	95.2	70.0	71.7	76.8	27.7	37.9	43.4
MoSo (Tan et al., 2024)	91.1	94.2	95.3	70.9	73.6	77.5	31.2	38.5	43.4
DP (Yang et al., 2023a)	90.8	93.8	94.9	-	73.1	77.2	-	-	-
Random*	93.0	94.5	94.8	74.4	75.3	77.3	41.5	42.8	43.1
UCB (Raju et al., 2021)	93.9	94.7	95.3	-	75.3	77.3	-	-	-
ϵ-Greedy (Raju et al., 2021)	94.1	94.9	95.2	-	74.8	76.4	-	-	-
InfoBatch (Qin et al., 2024)	94.7	95.1	95.6	76.5	78.1	78.2	42.2	43.2	43.8
Ours	**94.9**	**95.5**	**96.0**	**77.6**	**78.9**	**79.5**	**44.9**	**47.0**	**49.4**

ensuring that both underrepresented and semantically meaningful samples are utilized effectively.

We employ TrivialAugment as our augmenter, which is widely used and offers a computationally efficient augmentation strategy. During augmentation, only a single lightweight transformation per image is applied. This brings two key advantages: i). It introduces negligible computation overhead to the online training process, making it highly efficient. ii). Since each image undergoes only one transformation with a slight magnitude, the augmented samples remain within their original local feature space. This is well-suited for the objective of our dynamic data selection framework: filling intra-cluster gaps and enhancing decision boundaries within clusters. Thus, the consistency of selected samples is preserved while the data diversity in sparse regions is enhanced. Consequently, training using these augmented samples improves model performance. In addition, we provide the augmentation operation list used in Table **??** in the Appendix, which includes image transformations such as translation, rotation, equalization, etc.

Complexity analysis. The computational costs of our framework, when integrated into online training, are primarily associated with density estimation. Specifically, both querying and updating within the HNSW graph operates with a complexity of $\mathcal{O}(\log(n))$, where n is the total number of data points. Let T denote the total number of training epochs; then, the total cost is $\mathcal{O}(T * \log(n))$. Since $T \ll n$, the overall computational complexity remains $\mathcal{O}(\log(n))$, making our method scalable for large datasets. Furthermore, the data augmentation, as a standard pipeline in model training, introduces negligible overhead.

4. Experiment

4.1. Experiment Setup

Datasets and network architectures. In line with previous works (Tan et al., 2024; Xia et al., 2023b; Qin et al., 2024), we evaluate the effectiveness of our proposed method using widely adopted benchmark datasets, including CIFAR-10/100 (Krizhevsky et al., 2009), Tiny-ImageNet (Chrabaszcz et al., 2017), and ImageNet-1k (Deng et al., 2009). In addition, we evaluate the robustness of our method in noisy datasets. To further assess the generalization ability of our method, we extend the evaluation to more challenging datasets, such as ImageNet-A/O (Hendrycks et al., 2021b), ImageNet-Hard (Taesiri et al., 2024), and ImageNet-R (Hendrycks et al., 2021a). Additionally, we evaluate the generalization of our method across different deep architectures. Specifically, we conduct experiments using both ResNet-based, such as ResNet-18/50, and ViT-based models, such as ViT-B/L and Swin-Transformer, to demonstrate the robustness and scalability of our approach across diverse models.

Comparison with state-of-the-arts. We compare with our method both static and dynamic data selection methods, including 1) Random, 2) EL2N (Paul et al., 2021), 3) GraNd (Paul et al., 2021), 4) Herding (Welling, 2009), 5) Forgetting (Toneva et al., 2018), 6) Moderate-DS (Xia et al., 2023b), 7) Self-sup. prototypes (Sorscher et al., 2022), 8) MoSo (Tan et al., 2024), 9) DP (Yang et al., 2023a), 10) UCB (Raju et al., 2021), 11) ϵ-Greedy (Raju et al., 2021), 12) Glister (Killamsetty et al., 2021), and 13) InfoBatch (Qin et al., 2024).

Table 2. Results on ImageNet-1k with a 60% selection ratio using ResNet-50 on an 8-A100 server. Note that due to the high computational costs and device memory costs (Xia et al., 2023b), Glister and CG-Score are not reported. Some results are from (Qin et al., 2024). Time is the wall clock time; Overall (n*h) is the total GPU hour, where n is the node number.

Method	Herding	EL2N	GraNd	Forgetting	SSP	Moderate	MoSo	UCB	Infobatch	Glister	CG-Score	Ours	Whole Dataset
Acc. (%)	71.1	72.3	71.0	72.5	70.0	73.1	74.0	75.8	76.5	-	-	**76.9**	76.4
Time (h)	10.5	10.5	10.5	10.5	10.5	10.5	10.5	10.5	10.5	10.5	10.5	10.5	17.5
Overhead (h)	>17.5	>17.5	>17.5	>17.5	>24.0	>17.5	>17.5	0.03	0.0028	-	-	0.53	0.0
Overall (n*h)	>224.0	>224.0	>224.0	>224.0	>276.0	>224.0	>224.0	**84.0**	**84.0**	-	-	88.2	140.0

Table 3. Experimental results on Tiny-ImageNet with noisy and corrupted data using ResNet-50. The noisy ratio is 20%.

Method /	Noisy		Corrupted	
Selection Ratio (%)	20	30	20	30
Random	17.8	23.9	20.0	25.9
Herding	19.0	24.2	35.0	30.6
Moderate-DS	19.6	25.0	23.3	29.1
EL2N	13.9	18.6	18.6	24.4
GraNd	18.3	23.7	20.0	26.7
Forgetting	13.2	21.8	18.5	25.5
Self-sup. prototypes	15.1	21.0	20.2	26.9
CG-Score	8.4	15.3	16.4	24.4
Glister	21.6	25.5	21.2	22.0
MoSo	7.4	11.3	23.1	28.8
Random*	33.8	36.5	35.1	36.9
InfoBatch	34.9	37.1	35.1	38.1
Ours	**35.9**	**39.6**	**39.1**	**42.0**

Table 4. Experimental results on Tiny-ImageNet with data augmentation. The selection ratios are 30%, 50%, and 70%.

Whole Dataset	52.0		
Selection Ratio	30%	50%	70%
Random	29.8	37.2	42.2
Herding	31.6	39.2	45.6
EL2N	32.0	40.1	45.9
GraNd	32.2	40.5	46.2
Glister	33.1	42.2	46.5
Forgetting	27.2	36.2	44.2
Moderate-DS	33.8	41.5	46.6
Self-sup. proto.	33.4	41.1	46.6
MoSo	32.6	41.5	45.9
Random*	42.1	43.9	45.2
InfoBatch	43.2	45.9	48.3
Ours	**44.9**	**47.0**	**49.4**

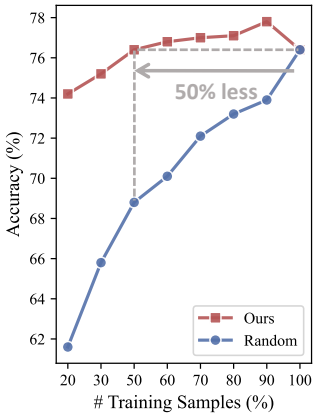


Figure 3. The performance on ImageNet-1k across various selection ratios with a 4-A100-GPU server.

Implementation details. To ensure consistency with prior work (Qin et al., 2024; Xia et al., 2023b), we follow similar experimental settings. Specifically, we use the OneCycle scheduler with the SGD/LARS optimizer for model training, a momentum of 0.9, a weight decay of 5e-4, and cosine annealing. We employ TrivialAugment (Müller & Hutter, 2021) in our framework. For fairness, we adopt the annealing and re-scaling techniques introduced in (Qin et al., 2024), which standardize the dynamic dataset pruning process across all methods compared. Moreover, we use InfoNCE loss to fine-tune adapters for 15 epochs on all datasets. Since InfoBatch uses soft pruning with a dynamic number of selected samples, we report its performance using the same number of forward passes as in our method.

4.2. Performance Comparison

As shown in Table 1, we evaluate the performance of our method by training ResNet-18 on CIFAR10/100 and ResNet-50 on Tiny-ImageNet across different selection ratios. Our method achieves comparable performance to models trained on the full dataset, even when only 50% of the data is used on CIFAR-10/100 and 30% on Tiny-ImageNet. In contrast, existing methods typically achieve lossless data selection with relatively higher selection ratios on these datasets, such as over 60% on CIFAR-10/100 and over 70% on Tiny-ImageNet.

Notably, our approach outperforms the other methods at the same selection ratios. On Tiny-ImageNet, a large-scale dataset, our method yields an average performance improvement of at least 2.7% while maintaining the same training costs. As the training data volume increases, this performance gap becomes even more pronounced, further highlighting the efficiency and effectiveness of our framework.

4.3. ImageNet-1k Results

Table 2 presents the evaluation results of our method on the ImageNet-1k dataset with a 60% selection ratio. Our approach outperforms the full dataset by achieving a nearly 40% training cost reduction, resulting in a reduction of up to 56 hours in training overhead with a 0.5% accuracy improvement. Meanwhile, since most static data selection methods require training surrogate models to determine the sample's influence throughout model training, the computation overheads are relatively much higher than ours. Thus, the results highlight that our method outperforms static data selection methods in both performance and computational efficiency. It also surpasses dynamic pruning methods in terms of final accuracy with comparable efficiency. These findings underscore the generality and competitiveness of our approach to large-scale datasets.

Further analysis of the performance across different selection ratios on ImageNet-1k is shown in Fig. 3. The results show that our method achieves lossless performance with only 50% of the training data. When using 20% of the data, performance drops by about 2% while nearly 80% of the training overhead is eliminated. Compared to random selection, which suffers a significant accuracy drop as

Table 5. Generalization of models trained with our method on ImageNet-Hard, ImageNet-A, ImageNet-R, and ImageNet-O. We report AUPR (%) for ImageNet-O and accuracy (%) on others. All models are ResNet-50.

Selection Ratio(%)	20	30	50	60	70	80	90	Whole Dataset
ImageNet-A	1.9$_{\downarrow 1.2}$	2.1$_{\downarrow 1.0}$	2.9$_{\downarrow 0.2}$	3.1$_{\uparrow 0.0}$	3.4$_{\uparrow 0.3}$	3.4$_{\uparrow 0.3}$	**3.5**$_{\uparrow \mathbf{0.4}}$	3.1
ImageNet-R	37.2$_{\uparrow 1.0}$	38.5$_{\uparrow 2.3}$	39.3$_{\uparrow 3.1}$	39.8$_{\uparrow 3.6}$	39.9$_{\uparrow 3.7}$	40.6$_{\uparrow 4.4}$	**41.0**$_{\uparrow \mathbf{4.8}}$	36.2
ImageNet-O	15.4$_{\uparrow 2.2}$	15.8$_{\uparrow 2.6}$	16.1$_{\uparrow 2.3}$	16.3$_{\uparrow 2.5}$	16.3$_{\uparrow 2.5}$	16.4$_{\uparrow 2.6}$	**16.5**$_{\uparrow \mathbf{2.7}}$	13.2
ImageNet-Hard	14.2$_{\downarrow 0.5}$	15.3$_{\uparrow 0.6}$	15.9$_{\uparrow 1.2}$	16.5$_{\uparrow 1.8}$	16.7$_{\uparrow 2.0}$	17.2$_{\uparrow 2.5}$	**17.5**$_{\uparrow \mathbf{2.8}}$	14.7

the selection ratios decrease, our method maintains robust performance even with reduced data. Similarly, most existing baseline methods typically require at least 60% of the training data to achieve similar lossless performance. This demonstrates that our framework further lowers the data requirement for maintaining full performance.

4.4. Robustness to Noisy Scenarios

In real-world scenarios, training data is often polluted by corrupted and mislabeled images (Xia et al., 2023a; Wang et al., 2018), which can significantly degrade model performance. Specifically, we simulate mislabeled data (noisy) by flipping a portion of the labels to incorrect ones using symmetric label noise. Meanwhile, we introduce five types of realistic distortions to simulate corrupted data, namely Gaussian noise, random occlusion, resolution variations, fog, and motion blur. The examples of corrupted data are shown in the Appendix. To evaluate the practical relevance of our data training framework in such noisy environments, we assess the robustness of our method compared to existing state-of-the-art methods.

As shown in Table 3, our approach consistently outperforms the compared methods, demonstrating superior robustness against both mislabeled and corrupted data. Specifically, our method achieves a 4% improvement over competing methods on corrupted datasets, even with a 20% noise ratio on Tiny-ImageNet. Our framework excels in these scenarios due to its ability to combine low-density sample selection with multimodal semantic alignment. While noisy data is typically sparse and low-density, our method's robust integration of multimodal semantics offers a powerful mechanism for mitigating noise and highlighting meaningful patterns in the data. This approach allows us to maintain high robustness without sacrificing data efficiency.

By prioritizing sparse, low-density samples and leveraging the corrective power of multimodal alignment, our method provides a reliable and efficient solution for robust deep learning in practical, noisy data environments, demonstrating its practical significance.

4.5. Effect of Data Augmentation on Model Performance

To further assess the effectiveness of our method, we compare its performance against several baseline methods using TrivialAugment for data augmentation, as shown in Table 4. Our method consistently outperforms the other approaches across various selection ratios.

Table 6. Experiment results on more advanced architectures, including ViT-B, ViT-L, and Swin-T on ImageNet-1k with a 4-A100 GPU server. Overhead represents GPU hours (h), and S_r refers to the selection ratio.

S_r(%)	50	60	70	80	90	Full Dataset
ViT-B	82.6	82.9	83.2	83.2	**83.3**	82.5
ViT-L	85.2	85.3	85.6	**85.7**	**85.7**	84.6
Swin-T	84.1	84.1	84.2	84.2	**84.3**	84.2

While data augmentation enhances the performance of all methods, our approach consistently achieves superior results at different selection ratios. This indicates that our method is not simply a straightforward combination of data augmentation and data selection. Instead, it effectively identifies the most beneficial samples for augmentation, leading to significant performance improvements. By selectively amplifying the impact of data augmentation, our method optimizes model performance, demonstrating its ability to leverage augmentation more effectively than other approaches.

4.6. Generalization on Hard Benchmarks

To evaluate the generalization capabilities of our proposed framework, we conduct experiments on challenging benchmark datasets, including ImageNet-Hard (Taesiri et al., 2024), ImageNet-R (Hendrycks et al., 2021a), and ImageNet-A/O (Hendrycks et al., 2021b). Specifically, we pre-train ResNet-50 models using data selected through our method across various selection ratios and then test their performance on these challenging benchmark datasets. Following standard evaluation settings, we report the area under the precision-recall curve (AUPR) for ImageNet-O and classification accuracy for the other datasets.

As shown in Table 5, our method maintains or even enhances generalization performance on these challenging datasets, despite using fewer training samples. The results demonstrate that reducing the dataset size with our framework does not compromise generalization ability. Meanwhile, as the selection ratios increase, our method achieves superior generalization compared to training on the full dataset.

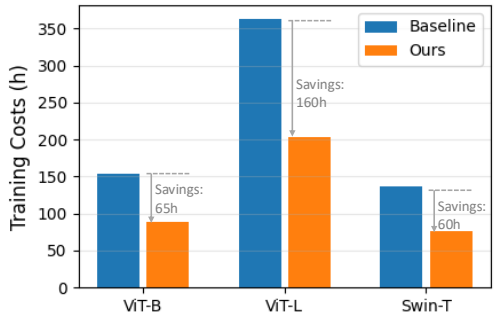
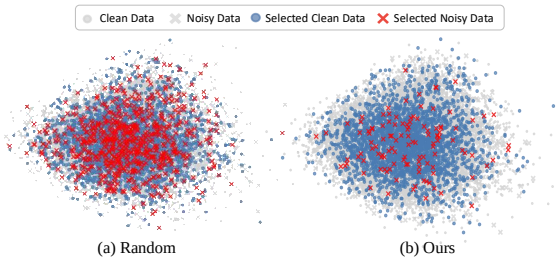


Figure 4. The overall cost savings achieved by our method on ViT-based architectures with lossless performance. The experiment is conducted on ImageNet-1k with a 4-A100-GPU server.



(a) Random (b) Ours

Figure 5. Visualization of the selection results on noisy Tiny-ImageNet with a 20% noise ratio. The selection ratio is 20%.

4.7. Generalization on Different Architectures

To evaluate the scalability of our proposed method, we conduct experiments on advanced architectures, including ViT-B, ViT-L (Dosovitskiy et al., 2020), and Swin-T (Liu et al., 2021). Specifically, we train these architectures using our framework across various selection ratios.

As shown in Table 6, our framework is architecture-agnostic, achieving robust generalization across these different models, even with reduced selection ratios. Notably, the lossless performance can be achieved with only 50% of the training data. The results underscore that our method generalizes on ResNet-based and Transformer-based architectures, all with reduced training costs.

Additionally, in Fig. 4, we present the practical training costs on these architectures and the lossless cost savings achieved by our framework. It can be seen that our proposed framework can significantly save hundreds of hours on large-scale architecture training.

4.8. Visualization of the Selection Robustness

In Fig. 1, we illustrate our selection results on clean datasets, showing that the selected data points mainly cluster around boundary regions among clusters. To better understand our selection effectiveness, in Fig. 5, we further illustrate our selection results on the noisy Tiny-ImageNet dataset with a 20% noise ratio. It can be seen that compared to the baseline, our method can effectively filter out noisy samples: the number of selected noisy points is minimized.

Table 7. Overheads of fine-tuning and feature embedding before model training on large-scale datasets with a 1-V100 GPU server.

Dataset	Fine-tuning	Embedding	Overall training
Tiny-ImageNet	0.39h	0.03h	21.0h
ImageNet-1k	1.25h	0.17h	84.0h

Table 8. Effect of density distribution, consistency distribution, and augmenter on Tiny-ImageNet using ResNet-50. We report test accuracy (%). The selection ratios (%) are 30%, 50%, and 70%.

p_ρ	p_{con}	aug.	30%	50%	70%
✓			39.0	40.7	42.5
	✓		42.0	45.6	45.8
		✓	41.6	45.9	48.3
	✓	✓	42.5	46.3	49.3
✓	✓		41.5	43.1	44.3
✓		✓	41.1	45.1	48.5
✓	✓	✓	**43.5**	**47.5**	**50.2**

4.9. Further Analysis of the Overheads

Although our method introduces negligible overheads into online training, our framework incorporates adapter fine-tuning and feature embedding via CLIP models and corresponding adapters before model training begins. As shown in Table 7, we analyze these pre-computation overheads of the adapter fine-tuning and feature embedding via CLIP models. It can be observed that these one-time overheads before model training are negligible compared to standard target model training. Once computed, no further computation is required during online training across selection ratios.

4.10. Ablation Study

Effect of different modules in our framework. In Table 8, we systemically evaluate the effectiveness of different components within our proposed framework on Tiny-ImageNet using ResNet-50 across various selection ratios.

When only the density distribution p_ρ is used, performance is lower, as low-density samples often include sparse and outlier data, which can introduce ambiguity into the training process. However, when both the density distribution p_ρ and consistency distribution p_{con} are combined, performance improves, demonstrating that incorporating semantic consistency helps mitigate the negative effects of density-based selection. Further performance gains are achieved by including the augmentation module, which boosts accuracy by a significant margin. This shows that augmentation plays a crucial role in improving performance, especially when the selected data points under p_ρ and p_{con} are more suited for augmentation, enhancing the model's generalization. The results indicate that removing any module from our framework leads to a substantial drop in performance.

5. Conclusion

This paper proposes a novel data training framework that unifies dynamic data selection and data augmentation for more enhanced model training acceleration. Unlike existing selection methods, our proposed approach identifies samples suitable for data augmentation. By combining this with augmentation, our framework can improve model generalization with reduced training costs. As a result, we can achieve lossless training acceleration with fewer data and enhanced generalization using the same volume of data. Extensive experiments demonstrate the effectiveness and efficiency of our method, especially in terms of generalization across large-scale datasets and more challenging scenarios.

References

Atienza, R. Improving model generalization by agreement of learned representations from data augmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 372–381, 2022.

Chrabaszcz, P., Loshchilov, I., and Hutter, F. A downsampled variant of imagenet as an alternative to the cifar datasets. *arXiv preprint arXiv:1707.08819*, Aug 2017.

Cubuk, E. D., Zoph, B., Mane, D., Vasudevan, V., and Le, Q. V. Autoaugment: Learning augmentation strategies from data. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 113–123, 2019.

Cubuk, E. D., Zoph, B., Shlens, J., and Le, Q. Randaugment: Practical automated data augmentation with a reduced search space. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds.), *Proc. Adv. Neural Inf. Process. Syst.*, volume 33, pp. 18613–18624, 2020.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

DeVries, T. and Taylor, G. W. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Du, J., Jiang, Y., Tan, V. Y., Zhou, J. T., and Li, H. Minimizing the accumulated trajectory error to improve dataset distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3749–3758, 2023.

Feldman, V. and Zhang, C. What neural networks memorize and why: Discovering the long tail via influence estimation. *Advances in Neural Information Processing Systems*, 33:2881–2891, 2020.

Gong, C., Wang, D., Li, M., Chandra, V., and Liu, Q. Keepaugment: A simple information-preserving data augmentation approach. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 1055–1064, 2021.

He, M., Yang, S., Huang, T., and Zhao, B. Large-scale dataset pruning with dynamic uncertainty. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7713–7722, 2024.

He, Z., Xie, L., Chen, X., Zhang, Y., Wang, Y., and Tian, Q. Data augmentation revisited: Rethinking the distribution gap between clean and augmented data. *arXiv preprint arXiv:1909.09148*, 2019.

Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8340–8349, 2021a.

Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., and Song, D. Natural adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15262–15271, 2021b.

Hong, T., Wang, Y., Sun, X., Lian, F., Kang, Z., and Ma, J. Gradsalmix: Gradient saliency-based mix for image data augmentation. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1799–1804. IEEE, 2023.

Hou, C., Zhang, J., and Zhou, T. When to learn what: Model-adaptive data augmentation curriculum. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1717–1728, 2023.

Iyer, R., Khargoankar, N., Bilmes, J., and Asanani, H. Submodular combinatorial information measures with applications in machine learning. In *Algorithmic Learning Theory*, pp. 722–754. PMLR, 2021.

Joshi, S. and Mirzasoleiman, B. Data-efficient contrastive self-supervised learning: Most beneficial examples for supervised learning contribute the least. In *International conference on machine learning*, pp. 15356–15370. PMLR, 2023.

Kang, M. and Kim, S. Guidedmixup: an efficient mixup strategy guided by saliency maps. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 1096–1104, 2023.

Killamsetty, K., Sivasubramanian, D., Ramakrishnan, G., and Iyer, R. Glister: Generalization based data subset selection for efficient and robust learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 8110–8118, 2021.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Lei, S. and Tao, D. A comprehensive survey to dataset distillation. *arXiv preprint arXiv:2301.05603*, 2023.

Lim, S., Kim, I., Kim, T., Kim, C., and Kim, S. Fast autoaugment. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Proc. Adv. Neural Inf. Process. Syst.*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/6add07cf50424b14fdf649da87843d01-Paper.pdf.

Lin, S., Zhang, Z., Li, X., and Chen, Z. Selectaugment: Hierarchical deterministic sample selection for data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 1604–1612, 2023.

Liu, S. and Wang, X. Mgdd: A meta generator for fast dataset distillation. *Advances in Neural Information Processing Systems*, 36, 2024.

Liu, T. Y. and Mirzasoleiman, B. Data-efficient augmentation for training neural networks. *Advances in Neural Information Processing Systems*, 35:5124–5136, 2022.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

Maharana, A., Yadav, P., and Bansal, M. D2 pruning: Message passing for balancing diversity and difficulty in data pruning. *arXiv preprint arXiv:2310.07931*, 2023.

Malkov, Y. A. and Yashunin, D. A. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE transactions on pattern analysis and machine intelligence*, 42(4):824–836, 2018.

Mirzasoleiman, B., Bilmes, J., and Leskovec, J. Coresets for data-efficient training of machine learning models. In *International Conference on Machine Learning*, pp. 6950–6960. PMLR, 2020a.

Mirzasoleiman, B., Bilmes, J., and Leskovec, J. Coresets for data-efficient training of machine learning models. In *International Conference on Machine Learning*, pp. 6950–6960. PMLR, 2020b.

Moreno-Barea, F. J., Jerez, J. M., and Franco, L. Improving classification accuracy using data augmentation on small data sets. *Expert Systems with Applications*, 161:113696, 2020.

Müller, S. G. and Hutter, F. Trivialaugment: Tuning-free yet state-of-the-art data augmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 774–782, 2021.

Nohyun, K., Choi, H., and Chung, H. W. Data valuation without training of a model. In *The Eleventh International Conference on Learning Representations*, 2023.

Northcutt, C. G., Athalye, A., and Mueller, J. Pervasive label errors in test sets destabilize machine learning benchmarks. *arXiv preprint arXiv:2103.14749*, 2021.

Paul, M., Ganguli, S., and Dziugaite, G. K. Deep learning on a data diet: Finding important examples early in training. *Advances in Neural Information Processing Systems*, 34: 20596–20607, 2021.

Poth, C., Sterz, H., Paul, I., Purkayastha, S., Engländer, L., Imhof, T., Vulić, I., Ruder, S., Gurevych, I., and Pfeiffer, J. Adapters: A unified library for parameter-efficient and modular transfer learning. *arXiv preprint arXiv:2311.11077*, 2023.

Qin, Z., Wang, K., Zheng, Z., Gu, J., Peng, X., xu Zhao Pan, Zhou, D., Shang, L., Sun, B., Xie, X., and You, Y. Infobatch: Lossless training speed up by unbiased dynamic data pruning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=C61sk5LsK6.

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

Raju, R. S., Daruwalla, K., and Lipasti, M. Accelerating deep learning with dynamic data pruning. *arXiv preprint arXiv:2111.12621*, 2021.

Ramalingam, S., Awasthi, P., and Kumar, S. A weighted k-center algorithm for data subset selection. *arXiv preprint arXiv:2312.10602*, 2023.

Singh, K. K. and Lee, Y. J. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, pp. 3544–3553. IEEE, Oct 2017.

Sorscher, B., Geirhos, R., Shekhar, S., Ganguli, S., and Morcos, A. S. Beyond neural scaling laws: beating power law scaling via data pruning. In Oh, A. H., Agarwal, A.,

Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022.

Sun, P., Shi, B., Yu, D., and Lin, T. On the diversity and realism of distilled dataset: An efficient dataset distillation paradigm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9390–9399, 2024.

Taesiri, M. R., Nguyen, G., Habchi, S., Bezemer, C.-P., and Nguyen, A. Imagenet-hard: The hardest images remaining from a study of the power of zoom and spatial biases in image classification. *Advances in Neural Information Processing Systems*, 36, 2024.

Tan, H., Wu, S., Du, F., Chen, Y., Wang, Z., Wang, F., and Qi, X. Data pruning via moving-one-sample-out. *Advances in Neural Information Processing Systems*, 36, 2024.

Toneva, M., Sordoni, A., Combes, R. T. d., Trischler, A., Bengio, Y., and Gordon, G. J. An empirical study of example forgetting during deep neural network learning. *arXiv preprint arXiv:1812.05159*, 2018.

Wang, Y., Liu, W., Ma, X., Bailey, J., Zha, H., Song, L., and Xia, S.-T. Iterative learning with open-set noisy labels. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8688–8696, 2018.

Welling, M. Herding dynamical weights to learn. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 1121–1128, 2009.

Xia, X., Han, B., Zhan, Y., Yu, J., Gong, M., Gong, C., and Liu, T. Combating noisy labels with sample selection by mining high-discrepancy examples. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1833–1843, 2023a.

Xia, X., Liu, J., Yu, J., Shen, X., Han, B., and Liu, T. Moderate coreset: A universal method of data selection for real-world data-efficient deep learning. In *The Eleventh International Conference on Learning Representations*, 2023b.

Xu, M., Yoon, S., Fuentes, A., and Park, D. S. A comprehensive survey of image augmentation techniques for deep learning. *Pattern Recognition*, 137:109347, 2023.

Yang, S., Li, J., Zhao, J., and Shen, F. Advmask: A sparse adversarial attack based data augmentation method for image classification, 2022a.

Yang, S., Xiao, W., Zhang, M., Guo, S., Zhao, J., and Shen, F. Image data augmentation for deep learning: A survey. *arXiv preprint arXiv:2204.08610*, 2022b.

Yang, S., Xie, Z., Peng, H., Xu, M., Sun, M., and Li, P. Dataset pruning: Reducing training data by examining generalization influence. In *International Conference on Learning Representations*, 2023a.

Yang, S., Guo, S., Zhao, J., and Shen, F. Investigating the effectiveness of data augmentation from similarity and diversity: An empirical study. *Pattern Recognition*, 148: 110204, 2024a.

Yang, S., Shen, F., and Zhao, J. Entaugment: Entropy-driven adaptive data augmentation framework for image classification. *arXiv preprint arXiv:2409.06290*, 2024b.

Yang, S., Ye, P., Ouyang, W., Zhou, D., and Shen, F. A clip-powered framework for robust and generalizable data selection. *arXiv preprint arXiv:2410.11215*, 2024c.

Yang, Y., Kang, H., and Mirzasoleiman, B. Towards sustainable learning: Coresets for data-efficient deep learning. In *International Conference on Machine Learning*, pp. 39314–39330. PMLR, 2023b.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=r1Ddp1-Rb.

Zhang, X., Du, J., Li, Y., Xie, W., and Zhou, J. T. Spanning training progress: Temporal dual-depth scoring (tdds) for enhanced dataset pruning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26223–26232, 2024.

Zheng, H., Liu, R., Lai, F., and Prakash, A. Coverage-centric coreset selection for high pruning rates. In *The Eleventh International Conference on Learning Representations*, 2023.

Zhong, Z., Zheng, L., Kang, G., Li, S., and Yang, Y. Random erasing data augmentation. In *Proc. AAAI*, volume 34, pp. 13001–13008, 2020.