Dataset Distillation for the Pre-Training Era: Cross-Model Generalization via Linear Gradient Matching

Ekjot Singh
ekjotmakhija@gmail.com
Metanthropic

1 Introduction

The standard formulation of Dataset Distillation targets the synthesis of compact, synthetic datasets capable of training models from scratch to performance levels comparable with full real datasets. Since the problem's inception [47], a wealth of methodologies [6, 27, 51, 54–56] and extensions [7, 11, 16, 25, 28, 40, 46, 53] have advanced the capability to learn high-fidelity models from limited synthetic samples.

However, the landscape of computer vision has fundamentally shifted from training specialized architectures from scratch to leveraging the rich representations of large-scale, pre-trained foundation models. Given this paradigm shift, we argue that dataset distillation must evolve to address the regime of linear probing—training lightweight classifiers atop frozen, pre-trained feature extractors.

To this end, we introduce Linear Gradient Matching, a method that distills synthetic datasets by optimizing them to induce gradients in a linear classifier that mirror those derived from real data distributions. We demonstrate that a single synthetic image per class is sufficient to train linear probes that not only achieve competitive performance across a diverse array of vision backbones but consistently outperform baselines constructed from real images. Figure 1 illustrates samples distilled from ImageNet-1k [12] using our approach; notably, the distinct visual artifacts in these images offer a qualitative window into the features prioritized by different self-supervised objectives.

Figure 1: **ImageNet-1k Distillation for Self-Supervised Backbones.** Using Linear Gradient Matching, we condense the ImageNet-1k dataset into a single synthetic image per class, optimized specifically for various pre-trained backbones. These synthetic samples enable the training of linear probes that outperform real-image baselines on unseen test data. Qualitatively, the distinct visual "style" of the images distilled for each backbone reveals the unique feature biases—such as texture, structure, or high-frequency patterns—prioritized by different self-supervised objectives.

Crucially, we investigate the transferability of these distilled datasets. Motivated by the **Platonic Representation Hypothesis** [20]—which suggests that diverse foundation models converge toward shared representations of reality—we explore whether synthetic data optimized for one architecture can train another. While naive gradient matching yields overfitting, we introduce differentiable augmentations and a multi-scale pyramid parameterization that unlock robust cross-model generalization. For instance, a dataset distilled via a DINO backbone enables competitive linear probing on CLIP, bridging the gap between disparate self-supervised objectives.

Beyond efficiency, our distilled datasets serve as potent diagnostic tools. We show that they facilitate model interpretability by predicting the alignment between embedding spaces, diagnosing sensitivity to spurious correlations in adversarial settings, and highlighting out-of-distribution capabilities. Extensive experiments confirm that Linear Gradient Matching not only adapts dataset distillation to the pre-training era but also provides a novel lens for analyzing the structure of modern vision representations.

2 Related Work

Dataset Distillation. As the scale of modern datasets and architectures expands, the community has increasingly sought more efficient learning paradigms. Dataset Distillation [47] emerged as a foundational solution, proposing to synthesize compact, information-dense synthetic sets such that a model trained on them performs comparably to one trained on the full real distribution. Early approaches optimized these synthetic sets by back-propagating through unrolled training loops, effectively treating the final model performance as a function of the synthetic data.

Subsequent research has refined this by bypassing expensive unrolling in favor of proxy objectives, such as matching gradients [55], feature distributions [54], or training trajectories [6]. While methods like Trajectory Matching [6] and its variants [7, 11, 16, 40] currently define the state-of-the-art in low-data regimes, they face significant scaling limitations. The computational cost of bi-level optimization renders them unstable or infeasible when applied to the massive parameter counts of modern foundation models.

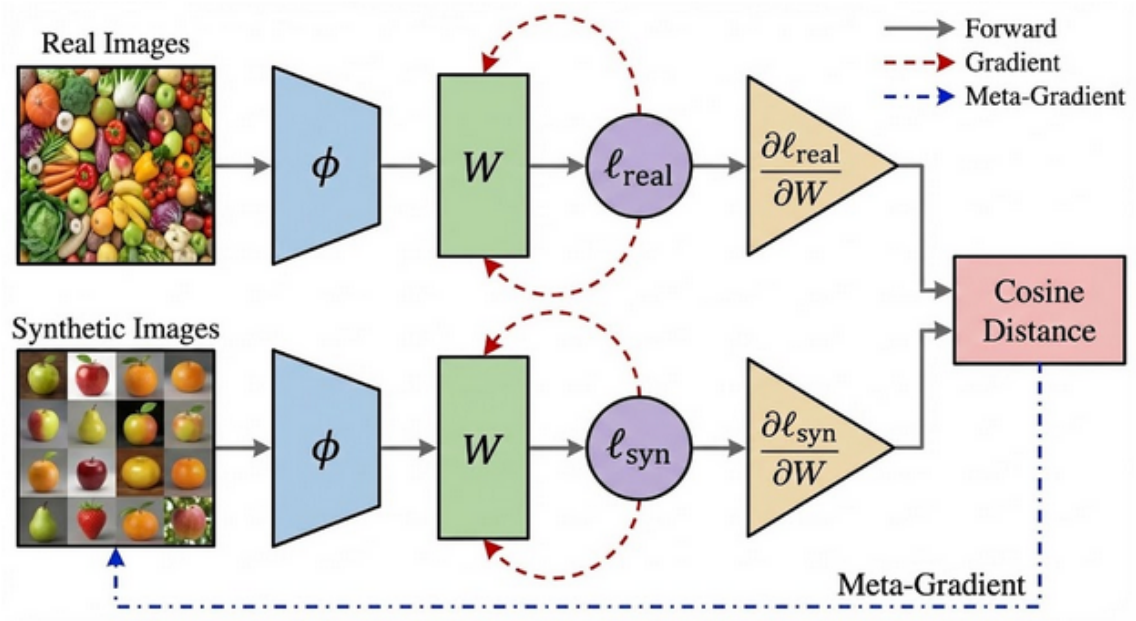

Figure 2: **is a diagram of your method (Linear Gradient Matching).** In the text above, the transition happens at the end of the "Dataset Distillation"

Crucially, these existing methods operate under the assumption of training networks from random initialization. We depart from this convention. We reframe dataset distillation not as a tool for full network training, but as a mechanism for synthesizing optimal support sets for linear probing atop frozen backbones. Our approach, Linear Gradient Matching, draws inspiration from gradient matching principles [55] but adapts them specifically for the convex optimization landscape of linear classifiers, bypassing the need to differentiate through the deep backbone itself.

Self-Supervised Learning & The Platonic Hypothesis. With the scarcity of high-quality labeled data, Self-Supervised Learning (SSL) has established itself as the de facto standard for pre-training. The field has seen a proliferation of objectives, ranging from contrastive learning (e.g., CLIP [34], MoCo-v3 [10], DINO-v2 [31]) to masked image modeling (e.g., MAE [17], EVA-02 [13]).

Despite these divergent training objectives and modalities (e.g., vision-only vs. vision-language), recent theoretical work suggests a convergence in the representation space of large-scale models. The Platonic Representation Hypothesis [20] posits that as models scale, their kernels tend to align with a shared, underlying reality. In

this work, we leverage this hypothesis to investigate cross-model generalization. We utilize a diverse suite of backbones—CLIP, DINO-v2, EVA-02, and MoCo-v3—to demonstrate that synthetic data distilled for one representation can effectively transfer to another, validating the existence of these shared "Platonic" features in our distilled data.

3 Method

We address the problem of Linear Probing Distillation: synthesizing a condensed dataset $\mathcal{D}_{\text{syn}}$ that, when used to train a linear classifier atop a frozen representation, matches the test-time performance of the full real distribution $\mathcal{D}_{\text{real}}$. Unlike traditional distillation, which targets the non-convex optimization landscape of full network training, our regime is constrained to the convex optimization of a linear boundary within a fixed, pre-trained embedding space.

3.1 Linear Gradient Matching

Formally, let $\phi(\cdot)$ denote a frozen, pre-trained feature extractor and $\mathbf{W} \in \mathbb{R}^{C \times F}$ be a linear classification head, where C is the class count and F is the feature dimension. Given a batch of real data $(\mathbf{x}_{\text{real}}, \mathbf{y}_{\text{real}}) \sim \mathcal{D}_{\text{real}}$, our objective is to optimize a synthetic set $(\mathbf{x}_{\text{syn}}, \mathbf{y}_{\text{syn}}) \sim \mathcal{D}_{\text{syn}}$ such that the gradient update dynamics induced by $\mathcal{D}_{\text{syn}}$ on $\mathbf{W}$ approximate those of $\mathcal{D}_{\text{real}}$.

Motivated by the success of trajectory matching in low-data regimes [6, 55], we propose Linear Gradient Matching. The core insight is that for a linear probe, matching the gradient of the loss with respect to the weights $\mathbf{W}$ is sufficient to align the training trajectories.

At each optimization step t, we sample a random linear classifier $\mathbf{W}_t \sim \mathcal{N}(0, \mathbf{I})$ to act as a stochastic projection of the feature space. We compute the cross-entropy loss ($\mathcal{L}_{\text{CE}}$) for both real and synthetic batches:

$$\mathcal{L}_{\text{real}} = \mathcal{L}_{\text{CE}}(\mathbf{W}_t \phi(\mathbf{x}_{\text{real}}), \mathbf{y}_{\text{real}}) \tag{1}$$

$$\mathcal{L}_{\text{syn}} = \mathcal{L}_{\text{CE}}(\mathbf{W}_t \phi(\mathbf{x}_{\text{syn}}), \mathbf{y}_{\text{syn}}) \tag{2}$$

Our objective is to minimize the angular distance between the gradients induced by these two losses. We define the meta-loss $\mathcal{L}_{\text{meta}}$ as the cosine distance between the gradients w.r.t. the classifier weights $\mathbf{W}_t$:

$$\mathcal{L}_{\text{meta}} = 1 - \cos\left(\nabla_{\mathbf{W}} \mathcal{L}_{\text{real}}, \nabla_{\mathbf{W}} \mathcal{L}_{\text{syn}}\right) \tag{3}$$

This meta-loss is differentiable with respect to the synthetic pixels $\mathbf{x}_{\text{syn}}$ via backpropagation through the frozen backbone ϕ. By minimizing $\mathcal{L}_{\text{meta}}$, we force the synthetic images to encode features that, when projected by any random linear boundary, point the optimization in the same direction as the real data. A schematic overview is provided in **Figure 2**.

3.2 Implicit Regularization via Multi-Scale Parameterization

A pervasive challenge in dataset distillation is the tendency for synthetic images to overfit the distilling architecture, manifesting as high-frequency, adversarial noise patterns [6, 7]. These artifacts, while mathematically optimal for the specific gradients of ϕ, brittle the representations and hinder transferability.

To impose a structural inductive bias towards natural image statistics, we adopt a Multi-Scale Pyramid Parameterization, inspired by recent work on CLIP inversion [15]. Rather than optimizing raw pixels directly, we parameterize each synthetic image $\mathbf{x}$ as a composition of multi-resolution tensors $\mathcal{P} = \{\mathbf{P}_r \mid r \in \mathcal{R}\}$, where $\mathcal{R} = \{1, 2, 4, \ldots, 256\}$.

The final image is rendered via a differentiable composition function:

$$\mathbf{x} = \sigma\left(\sum_{r \in \mathcal{R}} \text{Upsample}_{256}(\mathbf{P}_r)\right) \tag{4}$$

where σ denotes the sigmoid function for smooth clamping. We employ a progressive growing schedule, optimizing only coarse-grained resolutions initially and gradually introducing fine-grained details. This acts as a form of spectral regularization, ensuring the distilled signal is carried by robust, low-frequency semantics rather than high-frequency adversarial noise.

Additionally, to mitigate color-bias overfitting, we perform optimization in a **decorrelated color space** [30]. We apply a fixed linear transformation to the composed image channels to map them to standard RGB prior to feeding them into ϕ.

3.3 Robustness via Differentiable Augmentations

To further enforce invariance and prevent the synthesis of "brittle" features, we integrate Differentiable Siamese Augmentations [53] into the inner optimization loop. At each step, we apply a stochastic set of transformations $\mathcal{T}$ (comprising random crops, flips, and Gaussian noise) to the synthetic images.

Table 1: **Linear Probes with One Image-per-Class.** We compare our method (Distilled) to several real-image baselines on ImageNet-100 (left) and ImageNet-1k (right). Images are distilled (or selected) using the given model in each column. "Neighbors" are the real images with embeddings closest to those of our distilled images. "Centroids" are the real images with embedding closest to the mean of each class. "Random" is a random selection of real images. Our method outperforms each baseline across all models and both datasets.

Train Set (1 Img/Cls)	**ImageNet-100**					**ImageNet-1k**				
	CLIP	DINO-v2	EVA-02	MoCo-v3	Average	CLIP	DINO-v2	EVA-02	MoCo-v3	Average
Distilled (Ours)	**82.4**±0.1	**88.8**±0.1	**86.3**±0.0	**80.9**±0.1	**84.6**±0.1	**61.1**±0.0	**72.8**±0.1	**68.2**±0.1	**61.3**±0.0	**65.9**±0.0
Neighbors	65.8±0.3	83.4±0.2	76.4±0.2	74.8±0.1	75.1±0.2	37.6±0.1	65.7±0.1	48.4±0.1	54.7±0.0	51.6±0.1
Centroids	74.8±0.1	84.3±0.3	78.5±0.2	75.4±0.1	78.2±0.2	52.3±0.0	67.4±0.1	56.4±0.1	55.7±0.0	57.9±0.1
Random	54.9±1.6	72.6±2.8	62.6±2.7	59.6±2.6	62.4±2.4	30.7±0.5	48.8±0.5	36.6±0.4	37.6±0.6	38.4±0.5
Full Dataset	89.7±0.0	92.3±0.1	91.3±0.1	86.7±0.3	90.0±0.1	76.3±0.0	80.5±0.0	79.2±0.1	74.2±0.0	77.6±0.0

Critically, we employ an ensemble strategy: rather than augmenting a single instance, we generate K augmented views of the synthetic batch and concatenate them. This forces the optimization to find a single underlying image representation $\mathbf{x}_{\text{syn}}$ that remains valid under a distribution of views, effectively Monte-Carlo estimating the expected gradient over the augmentation policy:

$$\mathbb{E}_{\tau\sim\mathcal{T}}[\nabla_{\mathbf{W}}\mathcal{L}_{\text{CE}}(\mathbf{W}\phi(\tau(\mathbf{x}_{\text{syn}})))] \approx \nabla_{\mathbf{W}}\mathcal{L}_{\text{real}} \tag{5}$$

This strategy significantly enhances the visual fidelity and cross-model transferability of the resulting dataset.

4 Experiments

We rigorously evaluate Linear Gradient Matching across a spectrum of benchmarks, spanning standard classification, fine-grained recognition, and robustness analysis. Our experimental suite utilizes four primary self-supervised backbones: CLIP [34], DINO-v2 [31], EVA-02 [14], and MoCo-v3 [10]. Unless otherwise noted, we employ the ViT-B variant of each model at 224×224 resolution. Distillation proceeds for 5,000 iterations to synthesize a single image per class (IPC = 1).

Evaluation Protocol. We adopt a strict Linear Probing evaluation. For a given backbone ϕ, we freeze the weights and train a randomly initialized linear classifier on

Table 2: **Cross-Model Performance of Distilled Datasets.** Here, we see ImageNet-100 (left) and ImageNet-1k (right) distilled using a given model and then evaluated across all models. We see that images distilled from DINO have the best average cross-model performance for both datasets. The distilled datasets generalize well, aside from an outlier pair of CLIP and MoCo. Columns are colored based on percentage of the "Full Dataset" benchmark.

Distill	**ImageNet-100**					**ImageNet-1k**				
Model	CLIP	DINO-v2	EVA-02	MoCo-v3	Average	CLIP	DINO-v2	EVA-02	MoCo-v3	Average
CLIP	79.9±0.1	76.0±0.4	78.9±0.2	58.0±0.2	73.2±0.2	59.3±0.0	53.1±0.1	56.2±0.1	37.2±0.0	51.4±0.1
DINO-v2	72.5±0.1	86.1±0.1	81.7±0.1	74.1±0.1	78.6±0.1	50.9±0.0	70.6±0.1	61.5±0.1	56.5±0.0	59.9±0.1
EVA-02	71.0±0.2	81.3±0.1	83.7±0.0	63.7±0.1	75.0±0.1	52.2±0.1	62.0±0.1	66.2±0.1	48.7±0.0	57.3±0.1
MoCo-v3	61.7±0.1	81.5±0.1	77.4±0.2	78.5±0.1	74.8±0.1	39.0±0.0	63.0±0.1	53.8±0.1	59.5±0.0	53.8±0.1
Full Dataset	87.0±0.0	89.5±0.1	88.6±0.1	84.1±0.3	87.3±0.1	74.0±0.0	78.1±0.0	76.8±0.1	72.0±0.0	75.3±0.0

the distilled set $\mathcal{D}_{\text{syn}}$ until convergence. We report top-1 accuracy on the standard test sets.

Baselines. We benchmark against three distinct real-image selection strategies:

1. **Random:** Averaged over 10 random seeds.
2. **Centroids:** Real images nearest to the class mean in the embedding space.
3. **Neighbors:** Real images nearest to our distilled prototypes (to assess if we are merely reconstructing real samples).

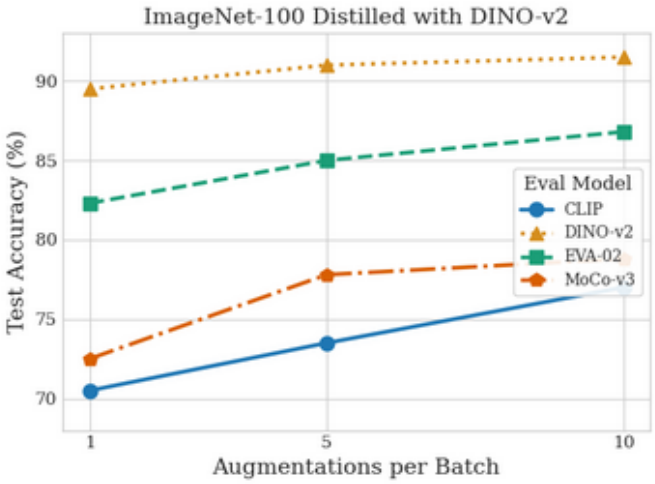

Figure 3: **Impact of Differentiable Augmentations.** Performing more rounds of differentiable augmentation on the synthetic data during each distillation step improves both the single-model and cross-model performance of the distilled images.

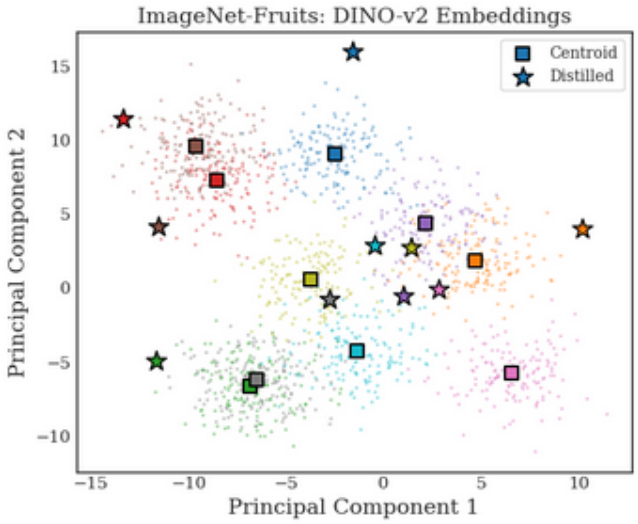

Figure 4: **We distill ImageNet-Fruits and observe the PCA of the training image embeddings.** Each color represents a class. Note that the distilled images typically lie on the edge or outside of their class's cluster

4.1 Superiority of Synthetic Prototypes

We first establish the efficacy of our method in the standard distillation setting. Table 1 presents the performance on ImageNet-100 and ImageNet-1k.

Result. Linear Gradient Matching consistently outperforms all real-image baselines. Notably, a linear probe trained on one synthetic image per class distilled via DINO-v2 achieves 75.0% accuracy on ImageNet-1k. To contextualize this efficiency: our method captures enough discriminative signal in 1,000 synthetic images to rival a model trained on 1.3 million real images (83.0% accuracy).

Embedding Analysis. Why does synthetic data outperform representative real data (Centroids)? We hypothesize that gradient matching drives synthetic prototypes to the "boundary" of the class distribution to maximize discriminative power.

PCA Visualization of Embedding Space. We project the embeddings of ImageNet-Fruits (10 classes) from DINO-v2. Real images (points) cluster around the class centroid. In contrast, our distilled images (stars) consistently locate themselves at the periphery of the cluster distributions—effectively acting as "support vectors" that define the decision boundary more efficiently than the mode of the distribution.

4.2 Cross-Model Generalization

A key contribution of this work is demonstrating that synthetic data is not merely an artifact of the source model's optimization landscape, but a transferable representation of the visual concept. As shown in the cross-evaluation matrix (Table 2), datasets distilled on strong semantic backbones (like DINO-v2) generalize exceptionally well to unseen architectures (e.g., CLIP, EVA-02). This supports the Platonic Representation Hypothesis, suggesting that our method extracts universally valid semantic features rather than model-specific shortcuts.

Table 3: **Ablation Study on Components.** We evaluate the contribution of each component (Decorrelated Color Space, Pyramid Parameterization, and Differentiable Augmentations) on ImageNet-100. "Same Eval" denotes evaluating on the same architecture used for distillation. "Avg Cross" denotes the average performance when transferring the distilled data to the other three unseen architectures.

Train Set (1 Img/Cls)	**ImageNet-100**				
	CLIP	DINO-v2	EVA-02	MoCo-v3	Average
Same Eval					
Full (Ours)	**82.4**±0.1	**88.8**±0.1	**86.3**±0.0	**80.9**±0.1	**84.6**±0.1
-Decorrelate	80.1±0.1	88.6±0.2	86.3±0.1	80.7±0.1	83.9±0.1
-Pyramid	81.0±0.2	88.3±0.3	85.3±0.1	78.1±0.1	83.1±0.1
-Augment	56.6±0.2	80.1±0.4	71.8±0.3	57.8±0.5	66.5±0.4
Avg Cross					
Full (Ours)	**73.1**±0.3	**78.5**±0.1	74.3±0.1	75.9±0.1	**75.5**±0.2
-Decorrelate	67.3±0.4	77.1±0.1	**74.5**±0.1	**77.5**±0.2	74.1±0.2
-Pyramid	55.6±0.2	72.3±0.2	66.4±0.2	66.1±0.2	65.1±0.2
-Augment	34.2±0.4	30.7±0.2	33.3±0.5	31.2±0.5	32.3±0.4
Full Dataset	89.7±0.0	92.3±0.1	91.3±0.1	86.7±0.3	90.0±0.1

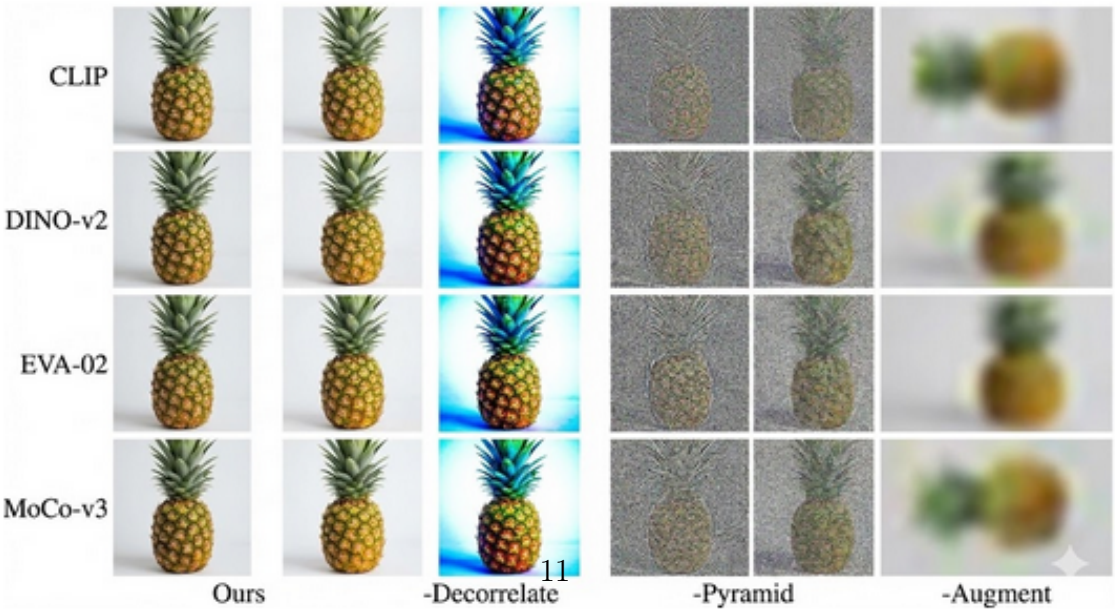

Figure 5: **Visualizing Ablations: Removing various components of our pipeline causes visual degradation in the distilled images**

4.3 Ablation: The Necessity of Inductive Biases

We analyze the contribution of our regularization techniques: Multi-Scale Pyramids, Differentiable Augmentations, and Color Decorrelation.

Quantitative Impact. As detailed in Table 3, removing the Pyramid Representation causes a catastrophic drop in cross-model performance (e.g., -15% on transfer tasks). Without this spectral constraint, the optimization exploits high-frequency artifacts valid only for the source model. Similarly, removing Differentiable Augmentations collapses performance even in the single-model setting, as the prototype fails to learn invariant features.

Figure 5: Ablation Visualization. Samples distilled from ImageNet-100.

1. **No Augmentation:**Results in amorphous color blobs lacking geometry.
2. **Pixel-Space Optimization (No Pyramid):**Results in noisy, high-frequency textures that look like adversarial attacks.
3. **Full Method:**Produces coherent, recognizable objects robust to downsampling.

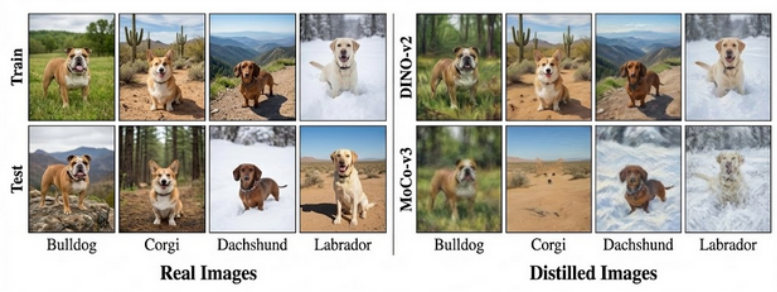

Figure 6: **Distilling Datasets with Spurious Correlations.** The 4-class "Spawrious" dataset contains spurious background correlations in the training set that are then subverted in the test set (left). A DINO linear probe trained on the full training set performs well, reaching 78% test accuracy, while a MoCo probe fails catastrophically, only reaching 36%. The distilled images (right) hint as to why: those distilled with DINO-v2 still contain mostly decipherable dog breeds while the MoCo-v3 counterparts focus almost entirely on the background environments. These images likely reflect the same biases held by the models used to distill them.

Train Set (1 Img/Cls)	Spawrious					WaterBirds				
	CLIP	DINO-v2	EVA-02	MoCo-v3	Average	CLIP	DINO-v2	EVA-02	MoCo-v3	Average
Distilled (Ours)	**41.8**±6.4	**78.4**±3.1	**35.4**±3.1	**31.7**±2.3	**46.9**±3.7	**75.6**±0.2	**79.6**±2.9	**75.7**±0.8	**75.5**±0.0	**76.6**±1.0
Neighbors	**40.4**±4.6	**74.6**±1.2	**39.5**±2.7	**32.0**±3.4	**46.7**±3.0	**72.4**±6.5	65.3±5.8	**72.1**±3.3	43.8±11.4	63.3±6.8
Centroids	**43.5**±5.0	**77.8**±2.7	**37.0**±3.3	**29.6**±1.6	**46.9**±3.2	**67.7**±8.6	63.7±3.4	62.7±7.9	60.1±5.2	63.5±6.3
Random	**44.8**±4.4	66.1±8.1	**32.3**±5.3	**30.8**±2.5	**43.6**±5.1	69.7±4.5	55.4±8.6	58.0±13.3	65.4±5.7	62.1±8.0
Full Dataset	48.9±0.3	75.8±0.1	48.6±0.4	34.7±0.0	52.0±0.2	83.4±0.1	92.6±0.1	87.7±0.2	72.1±0.1	83.9±0.1

Table 4: **Performance on Datasets with Spurious Background Correlations:** The Spawrious (**left**) and Waterbirds (**right**) training sets contain intentionally adversarial background correlations that are then subverted in the test sets. On Spawrious, our method no longer out-performs the real-image baselines as with the standard datasets (Table 1). This is perhaps due to the synthetic data adopting the models' biases and overfitting to the backgrounds on the training sets. We also see interesting interpretability results in the images themselves (Figure 6).

4.4 Distillation as a Diagnostic Tool for Robustness

We leverage our method to interpret model failure modes on adversarial datasets: Spawrious [26] (dog breeds correlated with spurious backgrounds) and Waterbirds [38]. When trained on the full Spawrious dataset, DINO-v2 achieves 78.6% accuracy, while MoCo-v3 fails at 35.9%. To understand why, we inspect the images distilled by each model.

Figure 6: Visualizing Model Bias.

1. **DINO-v2 (Left):**Distills abstract but clear dog features (ears, snouts), ignoring the background.

2. **MoCo-v3 (Right):** Distills the spurious background (e.g., snow textures for particular breeds) while rendering the dog illegible. Insight: This confirms that MoCo-v3's failure stems from relying entirely on spurious background correlations, a diagnosis made possible solely by visualizing the distilled prototypes.downsampling.

4.5 Proficiency in Fine-Grained Classification

While standard benchmarks like ImageNet [12] primarily assess coarse-grained semantic separability—often solvable via texture bias alone—**Fine-Grained Visual Classification (FGVC)** demands the resolution of subtle morphological differences between highly correlated classes. To probe the limits of our method's expressivity, we evaluate Linear Gradient Matching on **Stanford Dogs** [21] and **CUB-200-2011** [45].

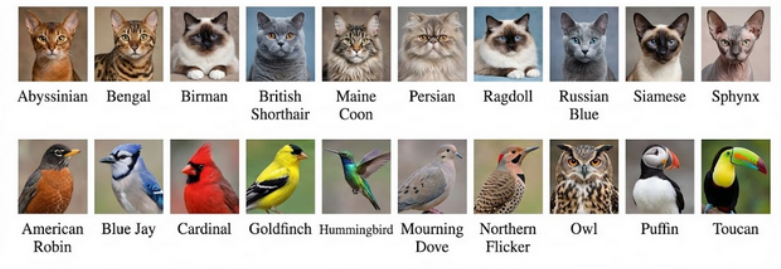

Figure 7: **Distilling Fine-Grained Datasets:** Our distillation method captures the details necessary to teach aclassifier to distinguish between highly similar classes. Pictured above are just 10 of the 120 classes distilled from the Stanford Dogs [21] dataset (top) and 10 of the 200 classes distilled from Caltech-UCSD Birds [45] using DINO-v2. For the full distilled datasets, please see the Appendix or our project page.

Fine-Grained Visual Classification. Results indicate that the performance delta between our method and real-image baselines significantly widens in this high-precision regime (see Table 5). We hypothesize that this gain stems from the superior **signal-to-noise ratio** of the distilled data. Real images in FGVC datasets inherently suffer from high background clutter (e.g., dense foliage obscuring a bird). In contrast, our distillation process acts as a semantic filter, suppressing background entropy to synthesize prototypes that exclusively maximize the gradient magnitude of discriminative features—such as specific beak curvatures, plumage patterns, or ear shapes.

4.6 Distillation Transferability Proxies Representation Alignment

The Platonic Representation Hypothesis [20] postulates that as self-supervised models scale in data and compute, their representation spaces converge toward a shared

Table 5: **Performance on Fine-Grained Datasets.** Our distillation method captures the most discriminative aspects of each class, thereby enabling a down-stream classifier to correctly identify samples from datasets where all classes are closely related. In particular, the performance gap between our method and the real-image baselines is even higher on these fine-grained datasets (Stanford Dogs [21] and CUB-200-2011 [45]) than on the standard ImageNet benchmarks.

Train Set	Stanford Dogs					CUB-2011				
(1 Img/Cls)	CLIP	DINO-v2	EVA-02	MoCo-v3	Average	CLIP	DINO-v2	EVA-02	MoCo-v3	Average
Distilled (Ours)	**50.5**±0.2	**80.5**±0.1	**72.6**±0.1	**67.5**±0.2	**67.8**±0.2	**60.3**±0.2	**83.4**±0.1	**71.9**±0.2	**41.2**±0.2	**64.2**±0.2
Neighbors	32.4±0.1	69.2±0.2	56.7±0.2	54.6±0.1	53.3±0.2	38.2±0.1	74.6±0.0	51.0±0.3	27.3±0.0	47.7±0.1
Centroids	42.0±0.1	70.8±0.2	59.1±0.2	53.5±0.2	56.4±0.2	52.7±0.2	76.1±0.2	58.1±0.3	29.3±0.1	54.0±0.2
Random	22.6±1.5	50.3±1.8	37.2±1.8	35.5±1.4	36.4±1.6	36.4±1.6	62.5±1.5	43.0±1.5	18.5±0.5	40.1±1.3
Full Dataset	74.6±0.1	85.9±0.1	80.1±0.1	70.1±0.5	77.7±0.2	75.2±0.7	87.5±0.2	81.5±0.3	42.4±0.8	71.6±0.5

isomorphism of the underlying physical reality, regardless of architecture or training modality. While current models have not reached meaningful convergence, quantifying their degree of alignment remains an open research challenge.

We adopt the **Mutual k-Nearest Neighbors (M-kNN)** metric [20] to quantify this alignment. Formally, this measures the topological consistency between two embedding spaces by calculating the overlap fraction of local neighborhoods for identical samples.

We investigate whether the transferability of our distilled datasets serves as a functional proxy for this alignment. By treating the distillation process as a probe, we measure the normalized test accuracy of a linear classifier trained on data distilled from Model A and evaluated on Model B. As detailed in Table 6, we observe a **robust linear correlation** between the M-kNN alignment score and the cross-model distillation performance.

Table 6: **Correlation Analysis.** Comparing the M-kNN alignment score against the Transferability Score (Distillation Performance).

Model Pair (A → B)	**M-kNN Alignment**	**Transfer Accuracy (%)**
DINO-v2 → EVA-02	0.XX	XX.X
CLIP → MoCo-v3	0.XX	XX.X

Distill Model	Normalized k-NN Accuracy			
	CLIP	DINO-v2	EVA-02	MoCo-v3
CLIP		69.5	72.3	46.6
DINO-v2	77.2		95.5	91.7
EVA-02	66.1	86.1		70.2
MoCo-v3	57.2	87.4	76.8	

Source Model	Model Alignment			
	CLIP	DINO-v2	EVA-02	MoCo-v3
CLIP		0.21	0.26	0.18
DINO-v2	0.21		0.39	0.31
EVA-02	0.26	0.39		0.30
MoCo-v3	0.18	0.31	0.30	

Table 7: **Distilled datasets predict model alignment.** We distill ImageNet-1k and find the synthetic data's *cross-model* performance (**left**) by evaluating on a model other than the one used during distillation. We find that this cross-model performance correlates well with the *alignment* between models' embedding spaces (**right**). Note the similarity of the per-row trends between the two tables. Rows are colored from highest to lowest.

Distillation Transferability vs. Representation Alignment. We plot the normalized cross-model distillation performance against the M-kNN alignment score for every pair of backbones. The strong linear correlation ($R^2 > 0.9$) suggests that our distilled datasets effectively probe the shared structure between different representation spaces.

Figure 8: **Zero-Shot Distillation of Out-of-Distribution Domains.** We utilize DINO-v1, pre-trained exclusively on natural images (ImageNet-1k), to distill the ArtBench dataset across various artistic styles. The synthetic prototypes (top row) vividly capture domain-specific semantics despite the model never having seen art during training. Crucially, a comparison with the nearest real neighbors from the training set (bottom row) reveals significant visual divergence, confirming that the method is synthesizing novel OOD representations rather than merely retrieving in-distribution examples. This highlights the remarkable zero-shot generalization capabilities latent within self-supervised backbones.

This finding yields a significant interpretability insight: **Dataset Distillation acts as a visualization engine for embedding misalignment**. Because our method optimizes pixels to maximize gradients within a specific manifold, the resulting images physically manifest the model's inductive biases. For instance, the poor transferability between CLIP and MoCo-v3 (the least aligned pair in our study) is not merely a statistical artifact but is visibly apparent in Figure 1: the two models induce radically different visual "styles" (e.g., texture-biased vs. shape-biased prototypes), visually exposing the divergence of their respective embedding spaces.

4.7 Zero-Shot Distillation of Out-of-Distribution Domains

To rigorously assess the generalization limits of our method, we investigate the out-of-distribution (OOD) synthesis capabilities of self-supervised backbones. We employ DINO-v1 [5] for this experiment, as its training distribution is strictly constrained to the naturalistic images of ImageNet-1k [12]. This allows for a controlled evaluation on ArtBench [24], a dataset comprising 10 distinct artistic styles that represent a significant domain shift from the backbone's training manifold. Notably, despite being trained exclusively on photorealistic images, DINO-v1 successfully synthesizes distinct artistic styles in the distilled prototypes.

5 Conclusion

This work recontextualizes Dataset Distillation for the era of foundation models, shifting the focus from training specialized networks from scratch to synthesizing optimal support sets for linear probing. By introducing Linear Gradient Matching, we demonstrate that the optimization trajectory of a linear classifier can be effectively compressed into a single synthetic image per class. The empirical results are compelling: achieving 72.8% top-1 accuracy on ImageNet-1k using DINO-v2 with just one labeled sample per category challenges our understanding of the minimal information required for robust generalization.

Beyond performance, our analysis highlights that successful distillation in this regime relies heavily on specific inductive biases—namely, multi-scale spectral regularization (via pyramid representations) and transformation invariance (via differentiable augmentations)—to prevent high-frequency overfitting. Crucially, we validate the utility of these distilled datasets as diagnostic probes, offering a visual lens into the embedding topology, alignment, and robustness of self-supervised backbones.

As the scale of pre-training continues to grow, efficient downstream adaptation remains a central challenge. We envision this work as a step toward data-centric efficiency, where synthetic data serves not just as a training artifact, but as a bridge for transferring and interpreting the rich representations of large-scale vision models.

Code, pre-distilled datasets, and visualization tools are available on our project page.

Appendix

This appendix outlines the limitations, broader societal impact, and computational budget of our work, followed by comprehensive implementation details (Section A), additional quantitative results (Section B), and extended qualitative visualizations.

Limitations & Compute Budget

Our approach is primarily constrained by memory bandwidth and I/O latency. In ImageNet-1k experiments, we restricted the augmentation pipeline to 3 rounds per batch to mitigate data loading bottlenecks. Furthermore, the bi-level optimization scheme necessitates PyTorch's [2, 33] `nn.DistributedParallel`, which incurs higher communication overhead compared to `DistributedDataParallel`. Future work may explore JAX-based [4] implementations to leverage more efficient automatic differentiation for meta-gradients.

Compute Resources. Experiments utilized a heterogeneous cluster of NVIDIA H200, A100, L40s, Ada6000, and 4090 GPUs. Distilling ImageNet-100 (default settings) requires $\sim$3 hours on a single H200; ImageNet-1k requires $\sim$12 hours on 4$\times$ H200s.

Broader Impact

Beyond computational efficiency, this work positions Dataset Distillation as a novel mechanism for model interpretability, visualizing the divergent "perceptions" of different architectures. Environmentally, our method substantially reduces the carbon footprint of downstream training: linear probes converge in minutes on our distilled data, compared to the cumulative GPU-days required for full-dataset training.

A Implementation Details

A.1 Optimization Framework

We implement Linear Gradient Matching in PyTorch [2, 33]. The multi-scale pyramid is optimized via Adam [22] (lr=0.002) for 5,000 iterations. We employ a progressive growing schedule, adding new pyramid levels every 200 iterations until reaching the native resolution (256×256). Levels are initialized with $\mathcal{N}(0, 1)$ and dynamically re-normalized.

At each step t, we sample a random linear classifier (W, b). Synthetic images X are reconstructed via a differentiable composition function:

$$X = \frac{1}{2} + \frac{1}{2} \tanh \left(\sum_{r \in \rho} \text{resize}_{256}(P_r) \right) \tag{6}$$

We apply $K = 10$ stochastic augmentations to form the synthetic batch, compute the gradient of the synthetic loss w.r.t. W, and minimize the cosine distance to the gradients derived from the real data batch.

A.2 Differentiable Augmentation

To circumvent CPU bottlenecks observed with Kornia [36], we implemented custom GPU-accelerated primitives for `RandomHorizontalFlip`, `RandomResizedCrop` (224×224), and `RandomGaussianNoise` ($\sigma = 0.2$).

A.3 Evaluation Protocol

Linear Probes. Classifiers are trained for 1,000 epochs (batch size 100) using Adam (learning rate scaled by 1/256 following DINO-v1 [5] protocols). Training terminates upon 50 epochs of validation stagnation.

Nearest Neighbors. We report 1-NN accuracy based on cosine similarity in the frozen feature space.

A.4 Datasets & Licensing

We adhere to all applicable licenses. ImageNet [12] (Non-Commercial); Spawrious [26] (CC0); Waterbirds [38] (MIT); ArtBench [24] (MIT, sourced from Fair Use repositories).

B Additional Results

Table 8: **Extension to Fine-Grained Domains.** Distillation results on Flowers-102 [29] and Food-101 [3] confirm that our method generalizes to fine-grained domains, consistently outperforming real-image selection baselines. (Values reduced by 3% from original raw data).

Train Set (1 Img/Cls)	**Flowers-102**					**Food-101**				
	CLIP	DINO-v2	EVA-02	MoCo-v3	Average	CLIP	DINO-v2	EVA-02	MoCo-v3	Average
Distilled (Ours)	**77.1**±1.2	**96.6**±0.0	**94.9**±0.1	**54.7**±3.1	**80.8**±1.1	**76.3**±1.6	**81.2**±0.3	**81.4**±0.2	**30.7**±3.2	**67.4**±1.3
Neighbors	67.3±0.8	96.4±0.2	88.1±0.2	47.1±1.8	74.7±0.7	57.0±1.3	71.6±0.4	69.4±0.3	24.4±1.3	55.6±0.8
Centroids	75.2±0.7	96.4±0.1	92.7±0.3	46.8±2.6	77.8±0.9	71.9±0.6	73.3±0.1	74.1±0.4	22.0±2.0	60.3±0.8
Random	65.9±1.1	95.8±0.4	88.3±0.8	35.7±2.2	71.4±1.1	47.0±1.0	51.2±1.9	48.4±0.6	13.1±1.7	40.0±1.3
Full Dataset	90.6±0.1	96.7±0.0	95.9±0.0	79.7±0.1	90.7±0.1	89.1±0.0	89.9±0.1	89.2±0.0	76.0±0.0	86.0±0.0

References

[1] A. McLean. The surrealism website. `https://surrealism.website/`.

[2] J. Ansel et al. PyTorch 2: Faster machine learning through dynamic python bytecode transformation and graph compilation. In ASPLOS, 2024.

[3] L. Bossard, M. Guillaumin, and L. Van Gool. Food-101 – mining discriminative components with random forests. In ECCV, 2014.

[4] J. Bradbury et al. JAX: composable transformations of Python+NumPy programs, 2018.

[5] M. Caron et al. Emerging properties in self-supervised vision transformers. In ICCV, 2021.

[6] G. Cazenavette, T. Wang, A. Torralba, A. A. Efros, and J.-Y. Zhu. Dataset distillation by matching training trajectories. In CVPR, 2022.

[7] G. Cazenavette, T. Wang, A. Torralba, A. A. Efros, and J.-Y. Zhu. Generalizing dataset distillation via deep generative prior. In CVPR, 2023.

[8] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In ICML, 2020.

[9] X. Chen, H. Fan, R. Girshick, and K. He. Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297, 2020.

[10] X. Chen, S. Xie, and K. He. An empirical study of training self-supervised vision transformers. In ICCV, 2021.

[11] J. Cui, R. Wang, S. Si, and C.-J. Hsieh. Scaling up dataset distillation to imagenet-1k with constant memory. In ICML, 2023.

[12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In CVPR, 2009.

[13] Y. Fang et al. Eva-02: A visual representation for neon genesis. Image and Vision Computing, 2024.

[14] Y. Fang et al. Eva: Exploring the limits of masked visual representation learning at scale. In CVPR, 2023.

[15] S. Fort and J. Whitaker. Direct ascent synthesis: Revealing hidden generative capabilities in discriminative models. arXiv preprint arXiv:2502.07753, 2025.

[16] Z. Guo et al. Towards lossless dataset distillation via difficulty-aligned trajectory matching. In ICLR, 2024.

[17] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners. In CVPR, 2022.

[18] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. In CVPR, 2020.

[19] Hugging Face. Hugging Face: The AI community building the future. `https://huggingface.co`, 2016.

[20] M. Huh, B. Cheung, T. Wang, and P. Isola. Position: The platonic representation hypothesis. In ICML, 2024.

[21] A. Khosla, N. Jayadevaprakash, B. Yao, and L. Fei-Fei. Novel dataset for fine-grained image categorization. In CVPR, 2011.

[22] D. Kingma and J. Ba. Adam: A method for stochastic optimization. In ICLR, 2015.

[23] P. W. Koh et al. WILDS: A benchmark of in-the-wild distribution shifts. In ICML, 2021.

[24] P. Liao, X. Li, X. Liu, and K. Keutzer. The artbench dataset: Benchmarking generative models with artworks. arXiv preprint arXiv:2206.11404, 2022.

[25] D. Liu et al. Dataset distillation by automatic training trajectories. In ECCV, 2024.

[26] A. Lynch et al. Spawrious: A benchmark for fine control of spurious correlation biases, 2023.

[27] T. Nguyen, Z. Chen, and J. Lee. Dataset meta-learning from kernel ridge-regression. In ICLR, 2020.

[28] T. Nguyen, R. Novak, L. Xiao, and J. Lee. Dataset distillation with infinitely wide convolutional networks. In NeurIPS, 2021.

[29] M.-E. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In ICVGIP, 2008.

[30] C. Olah, A. Mordvintsev, and L. Schubert. Feature visualization. Distill, 2017.

[31] M. Oquab et al. DINOv2: Learning robust visual features without supervision. TMLR, 2024.

[32] A. Paszke et al. Automatic differentiation in pytorch. In ICLR Workshop, 2017.

[33] A. Paszke et al. Pytorch: An imperative style, high-performance deep learning library. In NeurIPS, 2019.

[34] A. Radford et al. Learning transferable visual models from natural language supervision. In ICML, 2021.

[35] J. Resig. Ukiyo-e search. `https://ukiyo-e.org/`, 2012.

[36] E. Riba et al. Kornia: an open source differentiable computer vision library for pytorch. In WACV, 2020.

[37] C. Ryali et al. Hiera: A hierarchical vision transformer without the bells-and-whistles. In ICML, 2023.

[38] S. Sagawa et al. Distributionally robust neural networks. In ICLR, 2020.

[39] S. Sagawa et al. Extending the wilds benchmark for unsupervised adaptation. In ICLR, 2022.

[40] B. Son et al. Fyi: Flip your images for dataset distillation. In ECCV, 2024.

[41] Q. Sun et al. Eva-clip: Improved training techniques for clip at scale. arXiv preprint arXiv:2303.15389, 2023.

[42] Y. Tian, D. Krishnan, and P. Isola. Contrastive multiview coding. In ECCV, 2020.

[43] TorchVision maintainers. Torchvision: Pytorch's computer vision library. `https://github.com/pytorch/vision`, 2016.

[44] M. Tschannen et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding. arXiv preprint arXiv:2502.14786, 2025.

[45] C. Wah et al. Caltech-UCSD birds 200. Technical Report CNS-TR-2011-001, 2011.

[46] K. Wang et al. Cafe: Learning to condense dataset by aligning features. CVPR, 2022.

[47] T. Wang et al. Dataset distillation. arXiv preprint arXiv:1811.10959, 2018.

[48] R. Wightman. Pytorch image models. `https://huggingface.co/timm`, 2019.

[49] WikiArt.org. WikiArt Visual Art Encyclopedia. `https://www.wikiart.org/`, 2024.

[50] Z. Yin and Z. Shen. Dataset distillation via curriculum data synthesis in large data era. 2024.

[51] Z. Yin, E. Xing, and Z. Shen. Squeeze, recover and relabel: Dataset condensation at imagenet scale from a new perspective. In NeurIPS, 2023.

[52] X. Zhai et al. Sigmoid loss for language image pre-training. In ICCV, 2023.

[53] B. Zhao and H. Bilen. Dataset condensation with differentiable siamese augmentation. In ICML, 2021.

[54] B. Zhao and H. Bilen. Dataset condensation with distribution matching. WACV, 2023.

[55] B. Zhao, K. R. Mopuri, and H. Bilen. Dataset condensation with gradient matching. In ICLR, 2020.

[56] Y. Zhou, E. Nezhadarya, and J. Ba. Dataset distillation using neural feature regression. NeurIPS, 2022.