IDeaL: Data-Free Multi-Teacher Distillation via Improved Dead Leaves

Feyza Yavuz, Mert Bülent Sarıyıldız, and Diane Larlus

NAVER LABS Europe, France

Abstract. Multi-teacher distillation has emerged as a way to combine complementary teacher models into a single student model that exhibits the strengths of all its teachers. The student is trained to mimic the output of the teachers on a set of images, typically the union of the individual teacher's training sets, assuming this data is available. In this paper, we question that assumption and explore alternative options. We first study how far one can go when distilling from teachers fed with different types of noise. Then, we show that information contained in the teachers can be leveraged to tailor the noise for multi-teacher distillation: we propose a method that, thanks to decorrelation losses at both patch and image levels, generates teacher-specific, improved samples optimized for data-free distillation. Experiments show that our most effective samples, IDeaL, lead to strong students that successfully capture complementary information from the teachers, yielding surprisingly competitive results that substantially narrow the gap with students distilled from real images. Moreover, given a limited budget of 1K images for distillation, students distilled using our IDeaL samples match or surpass the performance of those distilled using a 1K-image subset of ImageNet.

Keywords: Multi-Teacher Distillation · Synthetic Data

1 Introduction

Vision Transformers (ViTs) [11] pretrained on large collections of web images have become the standard visual backbone, largely replacing CNNs. Today, it is common practice to use such pretrained visual encoders, which provide general-purpose feature representations to a wide range of downstream tasks. However, different models capture distinct visual characteristics, depending on their architecture, training process, or training data. Multi-teacher distillation has thus emerged as a popular way to reconcile these complementary views [19,46,48,52–54,59,69]; multiple teacher models jointly guide the learning of a single student, allowing it to inherit knowledge from all teachers at once.

However, multi-teacher distillation methods typically require large amounts of real images during training. These methods try to cover the input domain of the teachers by using their training data when available [48,52,53] or by using a large web-based dataset such as DataComp-1B [17] as in [19,46]. This reliance on real data poses practical challenges when the teachers' original training data is

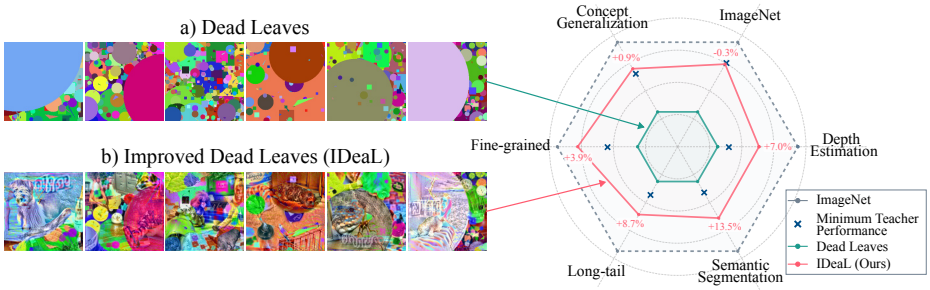

Fig. 1: In this paper, we distill complementary teachers into strong models, even without real data. **(Left)** Examples of the noisy samples that we use for distillation: i) procedurally generated synthetic Dead Leaves and ii) its improved version IDeaL. **(Right)** Performance on several downstream tasks of a student distilled from 4 teachers using 1M samples of either ImageNet, Dead Leaves, or our Improved Dead Leaves (IDeaL). We report the minimum performance obtained by the teachers on each task as a reference and denote the relative gains obtained by using IDeaL compared to the minimum teacher.

unavailable or cannot be redistributed due to legal restrictions, privacy concerns, or proprietary data policies. This issue is becoming relevant as foundation models are increasingly trained on proprietary data that is not publicly accessible [56,61]. As a result, obtaining suitable real-image datasets for distillation can be costly, impractical, or simply not possible.

In this paper we take an adversarial stance and ask: *how far can multi-teacher distillation go if not provided with any data?* Earlier work [53] has taken one step in this direction and replaced ImageNet [49] with realistic synthetic variants for each ImageNet class generated by Stable Diffusion [50]. We go further and pursue a completely *data-free* approach, where images are entirely replaced by procedurally-generated, synthetic surrogates. We start by analyzing how distillation performance degrades as the number of ImageNet samples decreases. We then replace real images with structured noise generated with procedural programs and compare several options already considered in data-free pretraining [3,27,38]. Among these, we find that Dead Leaves [3] consistently outperforms other alternatives, and shows encouraging transfer performance.

Building on these promising results, we propose a pixel-level optimization method that injects information from teacher attention layers into synthetic images, tailoring them for multi-teacher distillation. Starting from the best data-free alternative, Dead Leaves, we optimize these samples so that patches exhibit diverse representations within a single synthetic image while global representations differ across these images. To achieve this, we introduce decorrelation losses that encourage low cosine similarity between patches within an image and between image-level representations, yielding richer and more diverse inputs for distillation. Note that our synthetic images are not meant to recover the teachers' training sets, not even to look realistic. Instead, they are optimized to capture how the teachers process the data. This allows the student to learn from the teachers' complementary views, even without data.

To properly evaluate the effectiveness of our generation method and the resulting synthetic samples, we need a controlled distillation and evaluation setting. We choose to follow UNIC [53] as our main framework for multi-teacher distillation. In this framework, ImageNet is the dataset used by all teachers for training and distillation. That way, when we systematically replace it with synthetic alternatives, we can isolate the impact of our design choices on student performance. Moreover, the framework evaluates on a diverse set of classification and dense prediction tasks, providing a comprehensive overview of the distilled students' performance.

To summarize, our contributions are as follows. First, we study data-free multi-teacher distillation, systematically analyzing how distillation performance varies when reducing the amount of real data, and replacing it with procedurally generated synthetic datasets. Second, we propose a pixel-level optimization method, tailored to ViTs [11], that enriches synthetic images using complementary information provided by multiple teachers. Applied to the synthetic Dead Leaves samples, it produces a much more effective training set of Improved Dead Leaves (or IDeaL in short) for multi-teacher distillation.

Through extensive experiments in a controlled setting, we show that IDeaL samples yield strong students that close much of the performance gap with the students distilled on real images. Given a budget of only 1K images, they even match or surpass them. With 1M samples, they lead to strong students that are versatile enough to be competitive on all tasks, outperforming the minimum teacher performance on all considered tasks (except for ImageNet, the training set of the teachers), as shown in Fig. 1. To the best of our knowledge, this is the first work that fully replaces real images with synthetic surrogates in multi-teacher distillation while maintaining competitive performance across a large number of downstream tasks.

2 Related Work

In this section, we briefly review the literature relevant to our work. We first discuss recent knowledge distillation approaches and how those relate to our pipeline. Then we provide a short overview of methods that generate purely synthetic data tailored for different tasks or models.

Knowledge Distillation (KD) was introduced as a compression mechanism [5] to train a typically smaller student model using the outputs or intermediate features from a larger teacher model [21, 47, 70]. *Multi-teacher* distillation [19, 46, 48, 52–54, 59, 69] extends this idea by training a student to align with multiple teachers, each potentially pretrained under different objectives or data distributions, allowing it to build a unified representation suitable for many tasks. For effective distillation, prior work uses the same data for training the teachers and distilling to the student [48, 52, 53]. When teacher data is unavailable, [19, 46] resort to large web-based datasets [17] to cover the teachers' input domain. In both cases, these methods rely on large collections of real images. In contrast, our approach performs multi-teacher distillation *without any real images*: we

use only procedurally generated synthetic data, optimized via teacher attention features to be maximally informative for all teachers simultaneously.

Data-free knowledge distillation was introduced [33, 35, 39] to tackle the problem of teacher data not being available during distillation. A central idea is to exploit the teacher's knowledge to synthesize data that can be used for distillation, which often comes with strong assumptions about *e.g.* the availability of teacher's training data or class label information. For instance, [33] keeps a record of the teacher's activations extracted on the original training data. [7, 13, 35, 39, 43, 66] rely on class logits predicted by the teacher to train a generator that produces synthetic data, assuming the teacher was trained on a classification task, potentially requiring balanced data per class [7]. We avoid such assumptions, as we require neither teachers to be classification models nor any label information. Moreover, our approach does not suffer from the potential instability issues due to co-training a generator along with the student with an adversarial objective [43].

Other works decode the knowledge embedded in the teacher's internal statistics, such as BatchNorm (BN) [23], to guide data synthesis. In this context, [14, 22, 67, 68, 74] synthesize class-conditioned images by matching BatchNorm statistics stored in a pretrained CNN. [44] mitigates the activation covariate shift when providing Gaussian noise for distillation by keeping the teacher's BN modules in "training" mode. However, real data from the teacher's input domain is then needed at inference time to compute "running statistics" for the student's BN layers. Given that ViTs [11] use LayerNorm [1] instead of BN, it is not straightforward to apply these methods to modern ViT-based teachers.

Another line of work focuses on data-free model quantization of ViTs. PSAQ-ViT [31] is among the first works that propose a patch-similarity-based regularization for ViT-based models. It optimizes images from Gaussian noise, and subsequent works improve the inversion quality or efficiency [8, 22, 30, 45, 75, 76]. However, these ViT-targeted inversion methods are designed for a *single* model and for a different goal (*e.g.* for quantization) which requires only small-scale sample sets, making them difficult to generalize to multi-teacher settings where the student must learn from diverse teachers simultaneously. Our method differs from these works as it is designed from the ground up for the *multi-teacher* setting, jointly optimizing large-scale synthetic data across all teachers.

Procedural synthetic data. Recent work has explored formula-driven, procedurally generated data as alternatives to real images for pretraining [2, 3, 18, 24, 27, 38, 72], domain adaptation [58], and knowledge distillation [15, 16]. The premise is that such datasets enable truly data-free learning. They are not the output of generative models [12,50] trained to produce realistic-looking synthetic images from real ones. FractalDB [27] was introduced as formula-driven supervised learning and quickly extended [25, 26, 36–38, 57, 64, 65] to the point where fractals can replace real images for pretraining [36]. In parallel, Baradad *et al.* [3] showed that structured noise can yield competitive representations. One example is Dead Leaves, an image model generated using simple formulas. However, performance of models pretrained on such synthetic datasets does not scale with

the data size. Baradad *et al.* [2] later collected diverse procedural generation programs with better scaling properties, inspiring Frank *et al.* [15] to propose a data-free KD framework using procedurally rendered images. In follow-up work, Frank *et al.* [16] analyze what properties make surrogate datasets effective for KD. All these works use procedurally generated data *as-is* and focus on pre-training or single-teacher distillation. Our approach differs in two ways: first, we study the multi-teacher setting, where the student must simultaneously learn from teachers with heterogeneous training objectives; second, rather than using procedural images directly, we take Dead Leaves as a starting point and perform direct pixel-space optimization guided by multiple teachers' attention features, tailoring the synthetic data to the specific set of teachers.

3 Background

In this work, we train student models distilling from multiple pretrained teachers, where all models are ViTs [11]. First, we briefly introduce the ViT architecture and the multi-teacher distillation framework, and then discuss the data used for distillation, which is the main focus of this paper.

Vision transformers (ViTs) process input images by dividing them into a grid of P non-overlapping patches, which are then linearly projected into a sequence of patch tokens. A special `CLS` token is included in this sequence and serves as a global representation. Multiple layers of self-attention and feed-forward blocks are then applied to these tokens, allowing the model to learn complex relationships between different parts of the image. We denote the mapping from an input image $I \in \mathbb{R}^{H\times W\times 3}$ to the output features of a ViT encoder as $f(I)$. It produces a set of P patch features $\mathbf{F} \in \mathbb{R}^{P\times d}$ and the `CLS` feature $\mathbf{f} \in \mathbb{R}^{d}$, where d is the feature dimension of the model.

Multi-teacher distillation aims at distilling a given set of n pretrained teacher models $\{T_1, T_2, \ldots, T_n\}$ into a single student model S. Each teacher T_i is a ViT encoder whose mapping f_{T_i} produces: patch features $\mathbf{F}_{\mathrm{T}_i} \in \mathbb{R}^{P_{\mathrm{T}_i}\times d_{\mathrm{T}_i}}$ and `CLS` feature $\mathbf{f}_{\mathrm{T}_i} \in \mathbb{R}^{d_{\mathrm{T}_i}}$, where P_{T_i} and d_{T_i} are the number of patches and feature dimension of the i-th teacher, respectively. Similarly, the student model S is a ViT encoder with mapping f_{S} that produces patch features $\mathbf{F}_{\mathrm{S}} \in \mathbb{R}^{P_{\mathrm{S}}\times d_{\mathrm{S}}}$ and a `CLS` feature $\mathbf{f}_{\mathrm{S}} \in \mathbb{R}^{d_{\mathrm{S}}}$. For the sake of simplicity, we assume that the student and all teachers have the same number of patches $P_{\mathrm{S}} = P_{\mathrm{T}_i}$, and feature dimensions $d_{\mathrm{S}} = d_{\mathrm{T}_i}$ for all i. Some teachers may not have a `CLS` token; we replace it with the global average-pooled patch features, following [53]. The student is trained such that its features ($\mathbf{f}_{\mathrm{S}}$ and $\mathbf{F}_{\mathrm{S}}$) match every teacher's representation ($\mathbf{f}_{\mathrm{T}_i}$ and $\mathbf{F}_{\mathrm{T}_i}$). This is done by minimizing the distillation loss $\mathcal{L}^{\mathrm{Dist}} = \sum_{i=1}^{n} \mathcal{L}_i^{\mathrm{Dist}}$, which sums over teacher-specific distillation losses broadly defined as $\mathcal{L}_i^{\mathrm{Dist}} = \rho^{\mathrm{CLS}}(\mathbf{f}_{\mathrm{S}}, \mathbf{f}_{\mathrm{T}_i}) + \rho^{\mathrm{Patch}}(\mathbf{F}_{\mathrm{S}}, \mathbf{F}_{\mathrm{T}_i})$, where ρ^{CLS} and ρ^{Patch} are similarity measures between the student and teacher features, such as cosine similarity or smooth-ℓ_1 distance operating on the respective feature vectors. By optimizing on this combined objective over all teachers simultaneously, the student learns a unified

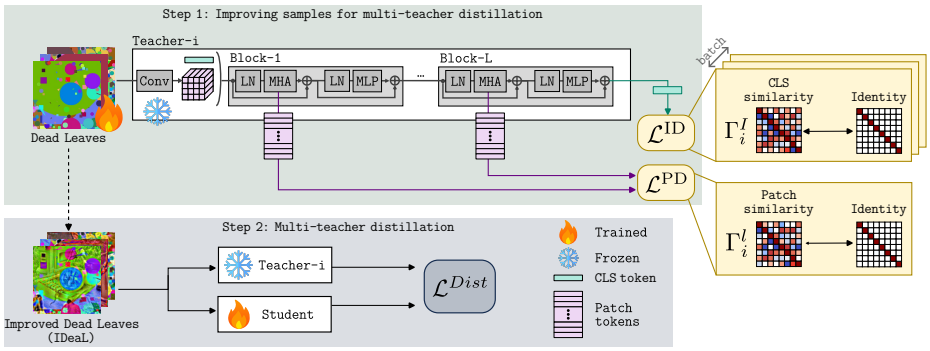

Fig. 2: Illustration of our 2-stage method for multi-teacher distillation without real data. **(Step 1)** Our method for improving input data tailored for multi-teacher distillation. Note that gradients of $\mathcal{L}^{\text{ID}}$ and $\mathcal{L}^{\text{PD}}$ are backpropagated through the frozen teacher encoders, which allows the synthetic images to be optimized to be maximally informative for all teachers simultaneously. **(Step 2)** Improved data is then used to train the student model distilling from all teachers, exactly as if it were real data.

representation that captures complementary knowledge encoded by the teachers. We closely follow the distillation framework from UNIC [53], see supplementary material for further details.

Distillation data. The loss $\mathcal{L}^{\text{Dist}}$ is computed over a set of training images. Existing multi-teacher distillation methods typically rely on large collections of real images to cover the input domain of the teachers. For instance, UNIC [53] uses the full ImageNet training set (1.28M images) while RADIO [19, 46] scales to 1B images and DUNE [52] combines more than 10 heterogeneous datasets. However, as mentioned in the introduction, the teachers' original training data may be unavailable due to licensing, privacy, or proprietary restrictions, so in this paper we study a drastically different approach where we assume access to no real data for distillation. We explore replacing real images entirely with Gaussian noise and procedurally generated synthetic data (such as FractalDB [27], and Dead Leaves [3]) and find that Dead Leaves consistently outperforms the other synthetic alternatives across all downstream tasks. Still, a gap remains between Dead Leaves and real data. In the next section, we propose a method to reduce this gap by *optimizing* synthetic images to be directly maximally informative of teachers, and hence better suited for multi-teacher distillation.

4 Optimizing images for multi-teacher distillation

Our goal is to train student S by distilling from multiple pretrained teachers $\{T_1, \ldots, T_n\}$ without relying on real data. To this end, we propose a method that generates synthetic images initialized from structured noise. These images are optimized to be maximally informative for all teachers jointly, so that distilling with such data yields a strong student. An observation from [31] is that ViT self-attention layers rely on rich patch representations to capture meaningful

spatial and semantic relationships within an image. However, when all patches look alike, as in Gaussian noise, the attention mechanism has little to work with. We therefore design generation objectives that encourage both *within-image* diversity, so that every patch carries distinct information, and *across-image* diversity, so that different generated samples cover complementary regions of teachers' feature spaces. These are defined as decorrelation losses operating on teacher attention features, explained in detail below.

Image optimization. Given an initial image $I \in \mathbb{R}^{H\times W\times 3}$, *e.g.* a sample from Dead Leaves, we treat its pixel values as learnable parameters. We then iteratively update these pixels by minimizing a generation loss $\mathcal{L}^{\mathrm{G}}$ (defined below), backpropagating through the frozen teacher encoders $f_{\mathrm{T}_1}, \ldots, f_{\mathrm{T}_n}$. Once optimized, the resulting synthetic images are stored and used as training data for distillation, *i.e.* to compute the distillation loss $\mathcal{L}^{\mathrm{Dist}}$ described in Sec. 3.

Patch decorrelation loss. Real images are quite complex, which leads to diverse patch representations in ViT layers. In contrast, unstructured noise yields nearly homogeneous patch representations, limiting the expressiveness of self-attention layers. To generate images with diverse internal structure, we encourage patch representations to be decorrelated across all teachers simultaneously.

For a given teacher T_i, we extract the attention output at each transformer layer $\ell \in \{1, \ldots, L\}$ and average across heads to obtain per-patch representations. We then compute the pairwise cosine similarity matrix $\Gamma_i^\ell \in \mathbb{R}^{P\times P}$, where entry (j, k) is the cosine similarity between the representations of the j-th and k-th patches. Inspired by self-supervised decorrelation objectives [71,73], we define the patch decorrelation loss as:

$$\mathcal{L}_i^{\mathrm{PD}} = \frac{1}{P^2} \sum_{\ell=1}^{L} \left\| \Gamma_i^\ell - \mathbf{I}_P \right\|_F^2, \tag{1}$$

where $\mathbf{I}_P \in \mathbb{R}^{P\times P}$ is the identity matrix and $\|\cdot\|_F$ is the Frobenius norm. Since the cosine similarity of a patch with itself is always one, the diagonal entries of Γ_i^ℓ are already at the target; minimizing Eq. (1) therefore drives all off-diagonal similarities to zero, making patches as dissimilar as possible within each image.

Image decorrelation loss. While $\mathcal{L}^{\mathrm{PD}}$ encourages within-image diversity, it does not prevent mode collapse, where different generated samples are highly similar to each other. To promote diversity across samples, we introduce a decorrelation loss on the global image representations. For each teacher T_i and for each image in the batch, we extract the final-layer `CLS` embedding (or spatially average-pooled patch embeddings) and compute the pairwise cosine similarity matrix on image level $\Gamma_i^{\mathrm{I}} \in \mathbb{R}^{B\times B}$, where B is the batch size. Then teacher-specific image decorrelation loss becomes:

$$\mathcal{L}_i^{\mathrm{ID}} = \frac{1}{B^2} \left\| \Gamma_i^{\mathrm{I}} - \mathbf{I}_B \right\|_F^2, \tag{2}$$

where $\mathbf{I}_B \in \mathbb{R}^{B\times B}$. Minimizing $\mathcal{L}_i^{\mathrm{ID}}$ ensures that each generated image has a distinct global representation as perceived by teacher T_i. If a teacher does not

use a learned `CLS` token, we use the global patch feature average as a surrogate, following the convention introduced in Sec. 3.

Total image generation loss. We combine both decorrelation losses across all teachers and add a regularization term $\mathcal{L}^{\mathrm{TV}}$, whose role is to encourage smoothness by penalizing the total variation of pixel values [68], leading to:

$$\mathcal{L}^{\mathrm{G}} = \alpha_1 \mathcal{L}^{\mathrm{TV}} + \sum_{i=1}^{n} \big(\alpha_2 \mathcal{L}_i^{\mathrm{PD}} + \alpha_3 \, \mathcal{L}_i^{\mathrm{ID}}\big), \tag{3}$$

where $\alpha_1, \alpha_2, \alpha_3$ are hyperparameters. By summing over all n teachers, generated images are optimized to be jointly informative across all teachers' feature spaces, which is essential to achieve strong multi-teacher distillation performance. Unlike [31, 68], our method does not assume any label or access to classification heads, making the method applicable to any combination of teachers, including self-supervised ones. The formulation relies only on matrix multiplications and Frobenius norms, which keeps backpropagation efficient and allows scaling to large synthetic datasets (up to a million samples in our experiments) across multiple teachers.

Training the student. The image generation and distillation stages are decoupled (Fig. 2): we first generate the full synthetic dataset by optimizing $\mathcal{L}^{\mathrm{G}}$, and then train the student by minimizing $\mathcal{L}^{\mathrm{Dist}}$ on the resulting images, exactly as if they were real data. This two-stage design means that the generation cost is paid once, after which the synthetic dataset can be reused across different student architectures or distillation configurations.

5 Results

5.1 Experimental protocol

Our framework consists of two main stages: synthetic sample generation and multi-teacher distillation. In the first stage, images are generated using our pixel-optimization loss, leveraging attention features from teachers to produce informative samples (Sec. 4). In the second stage, these synthetic samples are used to train the student model via multi-teacher distillation (Sec. 3).

We adopt the UNIC [53] multi-teacher distillation framework and evaluation protocol, with one critical modification: we replace the ImageNet dataset with different sets of synthetically generated samples.

Teachers and student. Following [53], we distill from four ViT-B/16 teachers for 100 epochs, each pretrained on ImageNet [10]: two self-supervised models, (i) DINO [6] and (ii) iBOT [78], and two supervised models, (iii) DeiT-3 [60] and (iv) a fine-tuned dBOT [32] optimized for ImageNet classification, namely dBOT-ft. All teachers use 768-dimensional features, 12 blocks, 224×224 input size and patch size 16. Importantly, all teachers are trained only on ImageNet [10], which ensures a controlled experimental setup, eliminating from our analysis

any bias that could come from the fact that teachers have been trained on different data. For instance, differences that we will see in generated synthetic images (Fig. 4) can be truly attributed to differences in the flavors of our method. The student is a ViT-B/16 model with the same architecture as the teachers.

Sample generation. We initialize the pixel optimization process with samples from Dead Leaves [3], resized to 224×224. All teachers are kept frozen, and only image pixels are trained. We use pools of 250 samples and optimize on randomly selected subsets of 40 samples. Each subset is used for 10 iterations before resampling, and each pool is optimized for a total of 4000 iterations. This increases interactions among examples within the samples while remaining computationally efficient. At each iteration, we apply random horizontal flips [31,68] to improve robustness. Teacher features are normalized at the end of the forward pass. $\mathcal{L}^{\text{PD}}$, $\mathcal{L}^{\text{ID}}$, and $\mathcal{L}^{\text{TV}}$ are weighted by 1, 1, and 0.05, respectively. We use the Adam [28] optimizer, with a learning rate set to 0.1. Unless stated otherwise, we generate 1K, 10K, 100K, and 1M samples for the main experiments (Tab. 1) and 10K samples per configuration for ablations.

Dataset for distillation. We compare the performance of student models distilled from the same teachers, but using one of these five different data sources: ImageNet [49], Gaussian noise, FractalDB [27], Dead Leaves [3], and improved Dead Leaves (IDeaL). For each source, we consider subsets of size 1K, 10K, 100K, and 1M samples, drawn uniformly at random, except for ImageNet, where we enforce a minimum of one sample per class. For each subset size, we report the average performance over three subsets obtained by different seeds.

Evaluation. We evaluate all models on the same downstream tasks as UNIC [53]. This includes image classification on the ImageNet validation set, transfer learning, semantic segmentation, and depth estimation. For transfer learning, we assess generalization on 15 image classification benchmarks: the 5 concept generalization levels of ImageNet-CoG [51], 8 fine-grained datasets (Aircraft [34], Cars196 [29], DTD [9], EuroSAT [20], Flowers [40], Pets [42], Food101 [4], and SUN397 [63]), and 2 long-tailed datasets (iNaturalist [62] 2018 and 2019). We report Top-1 accuracy for all classification tasks. For dense prediction, we report mIoU on ADE20K [77] and RMSE on NYUd [55], following the patch-level classification protocol of [41]. For all these tasks, we use linear probing on frozen encoder outputs.

5.2 From real images to data-free distillation

In this section, we first distill using real-image subsets of decreasing size, establishing an upper bound for the rest of our experiments. We then examine different types of procedural synthetic data and assess them in the context of multi-teacher distillation.

Distilling using real images, our oracle. Before considering data-free distillation, we assess the impact of reducing the distillation set when it consists

Table 1: Performance of teacher models and different students distilled from them, across different distillation data sources and sizes. We report Top-1 accuracy on ImageNet, average Top-1 accuracy on 15 classification tasks for transfer, mIoU on ADE20K for segmentation, and RMSE on NYUd for depth estimation. Best teacher performance and best student performance per distillation set size in **bold**. Student performance better than the min teacher is highlighted in beige. Relative improvements of IDeaL over Dead Leaves are shown in green. For students distilled on ImageNet, Dead Leaves and IDeaL subsets, we report mean over 3 different subsets.

	Model	Distillation data Source	Size	ImageNet Top-1(↑)	Transfer Top-1(↑)	Seg. mIoU(↑)	Depth RMSE(↓)
Teachers (all trained on ImageNet)							
1	DINO [6]			78.4	**72.4**	30.4	0.570
2	DeiT-3 [60]			83.6	68.3	32.3	0.589
3	dBOT-ft [32]			**84.0**	**72.4**	32.8	0.616
4	iBOT [78]			79.2	70.7	**36.6**	**0.524**
5	*Max teacher performance on each task*			84.0	72.4	36.6	0.524
6	*Min teacher performance on each task*			78.4	68.3	30.4	0.616
Students distilled using real ImageNet data - **oracle**							
7	UNIC [53]		1.28M	83.3	73.3	39.5	0.523
8		ImageNet	1M	**83.2**	**73.1**	**39.0**	**0.531**
9			100K	**81.7**	**71.6**	**37.8**	**0.542**
10			10K	**77.0**	66.5	**37.4**	**0.545**
11			1K	72.6	65.3	**34.2**	**0.566**
Students distilled using procedural synthetic data							
12		Gaussian Noise	1M	22.6	30.9	8.3	0.924
13			100K	22.7	31.3	8.2	0.915
14			10K	21.6	30.4	7.6	0.940
15			1K	21.1	29.4	7.2	0.965
16		Fractals	1M	32.7	37.3	10.9	0.932
17			100K	27.5	31.7	9.6	0.935
18			10K	27.1	31.7	9.4	0.925
19			1K	28.0	32.2	10.0	0.936
20		Dead Leaves	1M	66.9	65.5	29.1	0.632
21			100K	67.1	65.6	29.0	0.639
22			10K	66.6	65.1	29.1	0.655
23			1K	64.7	63.9	29.0	0.638
Students distilled using images optimized for the 4 teachers (**ours**)							
24		Improved Dead Leaves (IDeaL)	1M	78.2 ↑17%	70.6 ↑8%	34.5 ↑18%	0.573 ↑9%
25			100K	78.0 ↑16%	70.3 ↑7%	34.5 ↑19%	0.573 ↑10%
26			10K	**77.0** ↑16%	**69.9** ↑7%	34.5 ↑18%	0.592 ↑10%
27			1K	**74.1** ↑14%	**68.6** ↑7%	33.9 ↑17%	0.594 ↑7%

of real images. To this end, we create subsets of ImageNet of varying sizes (1K, 10K, 100K, and 1M) and distill with each.

We report the individual teacher results in `rows 1-4` of Tab. 1, with `rows 5-6` giving the maximum and minimum across teachers, which serve as reference

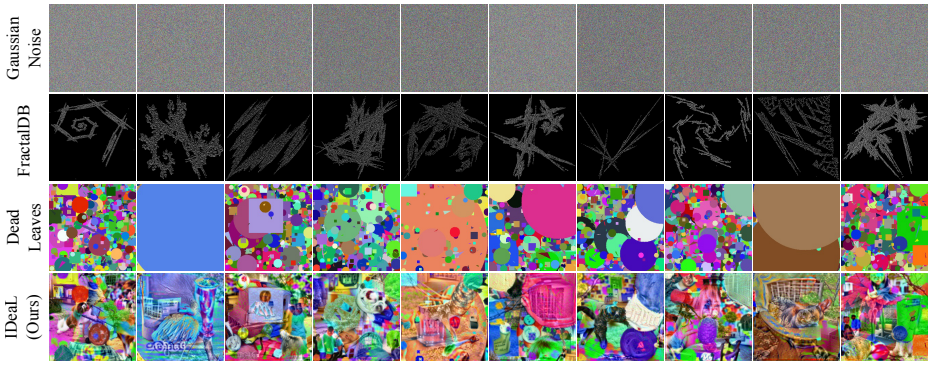

Fig. 3: Samples from the synthetic datasets considered in our experiments. From top to bottom: Gaussian noise, FractalDB, Dead Leaves, and our proposed Improved Dead Leaves (IDeaL). IDeaL optimizes Dead Leaves as described in Sec. 4, providing significantly richer information to the student while remaining fully data-free.

for assessing the students. The students distilled on subsets follow in `rows 8-11`, where `row 7` is the reproduced UNIC model [53], *i.e.* the student distilled on the full ImageNet dataset. We observe that reducing the distillation set from 1 million to 100K images barely affects results, showing that nearly 90% of ImageNet can be discarded with negligible impact. When further reducing to 10K images or fewer, the drop becomes far more drastic. This part of the table provides an upper bound on the results achievable for a given budget, and serves as our oracle for the rest of the discussion.

Distilling with procedural synthetic data. Next, we replace real images with procedural synthetic data during distillation. Unlike synthetic data obtained with a generator trained on real images [50], procedural synthetic image generation [27, 38] or structured noise [3] processes construct samples directly from predefined rules without learning from any real image dataset. While these samples have similar statistics to real images such as power spectrum [3], they do not contain any semantic content. Additionally, no privileged information from the original training dataset is used, such as the class information or any other dataset specific statistics. Inspired by [3], we consider Gaussian noise, FractalDB, and Dead Leaves as synthetic surrogate data for multi-teacher distillation (see examples in the first three rows of Fig. 3).

We compare the performance of students distilled with these different synthetic sets, `rows 12-23` of Tab. 1, to the students distilled with real data subsets, `rows 7-11`. We observe that students distilled from Gaussian noise yield poor results, regardless of the size of the distillation set. FractalDB performs better, but still lags behind Dead Leaves. Dead Leaves perform best among structured noise alternatives, and produce surprisingly strong transfer learning results. For instance, 100K Dead Leaves achieves 65.6 in transfer, on par with the 65.3 obtained with a 1K subset of ImageNet (our oracle using real images). Consequently, we use Dead Leaves as our starting point for the results presented in the next section.

Table 2: Performance of different student models distilled from the same four teachers, using varied flavors of Improved Dead Leaves as distillation data. Each distillation run is performed with a different set of 10K samples. Each set is obtained by optimizing with respect to a given set of teachers. The column `Joint Optim.` indicates whether the samples are optimized with respect to the teachers jointly (✓) or not. `Row 9` combines 4 sets of 2.5K samples, each optimized for one teacher only, whereas the 10K samples of `row 10` are obtained by optimizing with respect to the 4 teachers jointly. The best performance per column is in **bold**.

	Teachers used during sample optimization	Joint optim.	ImageNet Top-1(↑)	Transfer Top-1 (↑)	Seg. mIoU (↑)	Depth RMSE (↓)
Improved Dead Leaves optimized with 1 teacher						
1	DINO [6]	-	74.4	69.2	33.0	0.584
2	DeiT-3 [60]	-	74.3	68.9	32.2	0.597
3	iBOT [78]	-	74.9	69.1	33.9	0.578
4	dBOT-ft [32]	-	74.9	69.3	32.5	0.599
Improved Dead Leaves optimized with 2 teachers						
5	DINO, dBOT-ft	✓	75.9	**70.0**	34.0	0.605
6	DINO, DeiT-3	✓	75.5	69.3	33.1	0.589
7	iBOT, dBOT-ft	✓	76.0	69.6	34.2	0.575
8	iBOT, DeiT-3	✓	75.8	69.7	34.1	**0.573**
Improved Dead Leaves optimized with 4 teachers						
9	DINO, DeiT-3, iBOT, dBOT-ft	-	75.1	69.4	33.9	0.583
10	DINO, DeiT-3, iBOT, dBOT-ft	✓	**77.0**	69.9	**34.5**	0.592

5.3 Tailoring Dead Leaves for multi-teacher distillation

Dead Leaves yield strong results for a structured-noise-based distillation set; however, they are not yet competitive with real data. In this section, we enhance Dead Leaves using our method (Sec. 4) and report results in Tab. 1 (`rows 24-27`). We make the following key observations.

Observation 1: *Improved Dead Leaves (IDeaL) outperform the original Dead Leaves in all settings and for all tasks.* On the ImageNet classification task, the corresponding student improves for instance by 14% when using 1K samples (`row 23` vs. `row 27`), and by 17% (`row 20` vs. `row 24`) when using 1M samples. The largest improvements are observed for the semantic segmentation task, 18% relative gain for 1M setting. For transfer learning, the gains are smaller than for other tasks. To analyze this, we break down the transfer learning tasks into concept generalization, fine-grained, and long-tail classification. Across all three, IDeaL improves over Dead Leaves, by an overall 7%. The exception is long-tail tasks at larger subset sizes (100K and 1M), where IDeaL yields a 12% gain, while smaller subsets follow the trend. We refer the reader to the supplementary material for the per-sub-task results.

Observation 2: *Student models distilled on IDeaL outperform the minimum teacher performance for 3 tasks, even with as few as 1K samples.* Comparing

student and minimum teacher scores on each task reveals significant gains. Given that students trained with IDeaL do not have access to any information about the nature of the teachers or the data they were trained on, we find these results impressive. Even the students trained only with 1K IDeaL samples surpass the minimum teacher performance in transfer learning, semantic segmentation and depth estimation. Specifically, in transfer learning, we observe a +2.3 increase in Top-1 accuracy in the 1M setting, closing the gap with the maximum teacher performance. Moreover, in semantic segmentation, with a +4.1 gain in mIoU, the student outperforms all teachers except the top-performing one. Only on the ImageNet classification task are student models trained on IDeaL samples slightly below the minimum teacher performance (*i.e.* 78.2 *vs.* 78.4 when using 1M samples). This is not surprising, as all four teachers were trained on that dataset, making it inherently challenging to close this gap using only structured noise. However, we emphasize that, given access to complementary teacher models and no additional information, one can still distill a single versatile student with only 1K IDeaL samples that is more balanced across tasks than any individual teacher.

Observation 3: *In data scarce settings, students distilled with IDeaL perform on par with, or even better than students distilled on ImageNet subsets, in classification tasks.* Here, we focus on extreme cases with access to only 1K and 10K images as distillation datasets. Students distilled on IDeaL subsets outperform those distilled on ImageNet subsets by +3.3 and +3.4 Top-1 accuracy on the 1K and 10K sets, respectively, for transfer learning. On ImageNet classification, the 1K IDeaL student outperforms the 1K ImageNet student by +1.5 accuracy. Moreover, the 10K IDeaL students match the 10K ImageNet ones.

These results demonstrate that the optimized samples are highly informative, effectively encoding information from the teachers' representations, and enabling students to capture more of their knowledge. However, this does not generalize to larger subsets (100K or 1M). Similar to the scaling trend noted in [3], procedural synthetic data, including IDeaL, does not scale as well as ImageNet, particularly for dense tasks. We provide a scaling analysis in the supplementary material.

5.4 Distilling to different architectures

To ensure a fair evaluation of IDeaL, all experiments until here were conducted within the UNIC [53] framework without modification, using a ViT-B/16 student and ViT-B/16 teachers. Since IDeaL samples are optimized leveraging information from the teachers only, they do not depend on the student architecture and can in principle be used to distill these teachers into any student. To verify this we distill the same four ViT-B/16 teachers into a smaller ViT-S/16 student, using the same IDeaL samples used to produce the results reported in Tab. 1. In Tab. 3, we report the performances for our smallest (1K) and largest (1M) subset sizes. We find that our observations 1 and 3 also hold in this setup. IDeaL outperforms Dead Leaves across all tasks. Moreover, for observation 3, the gains at 1K samples are even more pronounced than those observed for ViT-B/16

Table 3: Different student architecture (ViT-S/16). Distillation of the four ViT-B/16 teachers into a ViT-S/16 student under the UNIC setup. Metrics as in Tab. 1, for the 1K and 1M subset sizes; ImageNet is the real-data upper bound. Relative improvements of IDeaL over Dead Leaves are shown in green.

Distillation data	Size	ImageNet Top-1(↑)	Transfer Top-1(↑)	Seg. mIoU(↑)	Depth RMSE(↓)
ImageNet (Oracle)	1M	**80.8**	**69.6**	**36.6**	**0.572**
Dead Leaves	1M	59.2	60.1	24.2	0.673
IDeaL (Ours)	1M	72.6 ↑23%	66.3 ↑10%	30.6 ↑26%	0.634 ↑6%
ImageNet (Oracle)	1K	53.7	49.6	27.6	0.640
Dead Leaves	1K	55.8	57.9	23.3	0.690
IDeaL (Ours)	1K	**66.8** ↑20%	**63.4** ↑9%	**29.6** ↑27%	**0.632** ↑8%

Table 4: Ablation study on our proposed loss components during optimizing Dead Leaves samples. The first row corresponds to the original Dead Leaves whereas the other rows correspond to optimized ones for different loss combinations. We generate 10K samples for each configuration. $\mathcal{L}^{\text{PD}}$ and/or $\mathcal{L}^{\text{ID}}$ are always used together with $\mathcal{L}^{\text{TV}}$.

	$\mathcal{L}^{\text{ID}}$	$\mathcal{L}^{\text{PD}}$	ImageNet Top-1(↑)	Transfer Top-1(↑)	Seg. mIoU(↑)	Depth RMSE(↓)
1	-	-	66.6	65.1	29.1	0.655
2	✓	-	73.4	68.2	32.4	0.663
3	-	✓	75.4	69.1	33.4	0.595
4	✓	✓	**77.0**	**69.9**	**34.5**	**0.592**

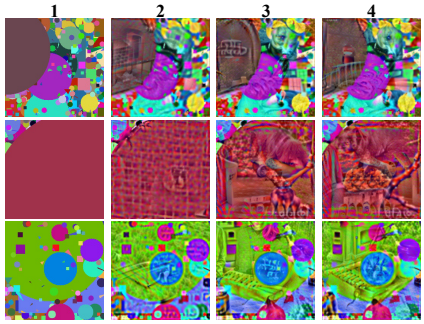

Fig. 4: Samples generated with different losses. Each loss contributes distinct structural properties.

in Tab. 1: IDeaL outperforms Dead Leaves by 20% on ImageNet classification and 27% on semantic segmentation, while also surpassing ImageNet on all tasks. Consistent with the scaling limitations of synthetic data, at larger subset sizes (1M), a performance gap with ImageNet images remains.

The exact set of teachers used during IDeaL generation matters. A question may arise: *Would optimizing IDeaL for only a subset of the teachers lead to data that is equally effective for distillation?* First, to investigate the effect of the set of teachers on the generated samples, in Tab. 2 we report results based on samples optimized using only one teacher (`rows 1-4`), two teachers (`rows 5-8`), and four teachers (`row 10`). For all settings, combining more teachers boosts performance, and optimizing with all teachers produces the strongest results. We also compare a different way of using all teachers. In `row 9`, we combine an equal number of samples (2.5K) from each set of samples optimized for one teacher independently. This set is then compared to the original IDeaL (`row 10`), which is jointly optimized with all teachers. The clear difference between these two rows shows that our method effectively projects information from multiple teachers

simultaneously. As an additional remark, samples generated using iBOT, which performs best on dense tasks, lead to a student that also performs well on these tasks, compared to students trained on samples generated with other teachers.

Ablation on loss components. To analyze the impact of the components of our sample optimization loss Eq. (3), we provide both qualitative (Fig. 4) and quantitative (Tab. 4) results. Our reference point (`row 1`) is distilled using the original Dead Leaves dataset. We compare this row to IDeaL optimized using only $\mathcal{L}^{\text{ID}}$ (`row 2`), only $\mathcal{L}^{\text{PD}}$ (`row 3`), and using both (`row 4`), which corresponds to our standard IDeaL setting. Note that the $\mathcal{L}^{\text{TV}}$ term from Eq. (3) is always used for regularization.

Results show that $\mathcal{L}^{\text{PD}}$ is the main source of performance gains, while $\mathcal{L}^{\text{ID}}$ provides a complementary boost. Using $\mathcal{L}^{\text{ID}}$ alone is insufficient, particularly for dense tasks, and the best results are obtained when both losses are combined. Additionally, the qualitative effect of each loss can be seen in Fig. 4, illustrating how each component contributes differently to the structure and properties of the generated images. Repeating this ablation with Gaussian noise instead of Dead Leaves as initialization yields the same conclusions; see the supplementary material for details.

6 Conclusion

In this paper, we investigated the notion of data-free multi-teacher distillation. We first analyzed the effectiveness of structured noise in that context. Next, to close the performance gap between distilling with structured noise and with real images, we proposed a pixel-level optimization objective that improves structured noise in a way that effectively captures information from all teachers. We observed that these synthetic samples, IDeaL, when used as distillation sets, significantly boost student performance across all tasks compared to a distillation set of non-optimized samples. Our optimization objectives are simple, require no teacher or dataset-specific information, and are truly data-free. As a result, they can potentially be applied to any other type of synthetic data, and extended to any kind of teachers or downstream tasks.

References

1. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv:1607.06450 (2016)
2. Baradad, M., Chen, C.F., Wulff, J., Wang, T., Feris, R., Torralba, A., Isola, P.: Procedural image programs for representation learning. In: Proc. NeurIPS (2022)
3. Baradad, M., Wulff, J., Wang, T., Isola, P., Torralba, A.: Learning to see by looking at noise. In: Proc. NeurIPS (2021)
4. Bossard, L., Guillaumin, M., Van Gool, L.: Food-101 – Mining discriminative components with random forests. In: Proc. ECCV (2014)
5. Buciluǎ, C., Caruana, R., Niculescu-Mizil, A.: Model compression. In: Proc. SIGKDD (2006)

6. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proc. ICCV (2021)
7. Chen, H., Wang, Y., Xu, C., Yang, Z., Liu, C., Shi, B., Xu, C., Xu, C., Tian, Q.: Data-free learning of student networks. In: Proc. ICCV (2019)
8. Choi, K., Lee, H., Kwon, D., Park, S., Kim, K., Park, N., Choi, J., Lee, J.: Mimiq: Low-bit data-free quantization of vision transformers with encouraging inter-head attention similarity. Proc. AAAI (2025)
9. Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., Vedaldi, A.: Describing textures in the wild. In: Proc. CVPR (2014)
10. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: Proc. CVPR (2009)
11. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: Proc. ICLR (2021)
12. Fan, L., Chen, K., Krishnan, D., Katabi, D., Isola, P., Tian, Y.: Scaling laws of synthetic images for model training ... for now. In: Proc. CVPR (2024)
13. Fang, G., Song, J., Shen, C., Wang, X., Chen, D., Song, M.: Data-free adversarial distillation. arXiv:1912.11006 (2020)
14. Fang, G., Song, J., Wang, X., Shen, C., Wang, X., Song, M.: Contrastive model inversion for data-free knowledge distillation. In: Proc. IJCAI (2021)
15. Frank, L., Davis, J.: Data-free knowledge distillation using adversarially perturbed opengl shader images. arXiv:2310.13782 (2023)
16. Frank, L., Davis, J.: What makes a good dataset for knowledge distillation? In: Proc. CVPR (2025)
17. Gadre, S.Y., Ilharco, G., Fang, A., Hayase, J., Smyrnis, G., Nguyen, T., Marten, R., Wortsman, M., Ghosh, D., Zhang, J., Orgad, E., Entezari, R., Daras, G., Pratt, S., Ramanujan, V., Bitton, Y., Marathe, K., Mussmann, S., Vencu, R., Cherti, M., Krishna, R., Koh, P.W., Saukh, O., Ratner, A., Song, S., Hajishirzi, H., Farhadi, A., Beaumont, R., Oh, S., Dimakis, A., Jitsev, J., Carmon, Y., Shankar, V., Schmidt, L.: Datacomp: in search of the next generation of multimodal datasets. In: Proc. NeurIPS (2023)
18. Guss, W.H., Houghton, B., Topin, N., Wang, P., Codel, C., Veloso, M., Salakhutdinov, R.: Minerl: A large-scale dataset of minecraft demonstrations. In: Proc. IJCAI (2019)
19. Heinrich, G., Ranzinger, M., Yin, H., Lu, Y., Kautz, J., Tao, A., Catanzaro, B., Molchanov, P.: Radiov2.5: Improved baselines for agglomerative vision foundation models. In: Proc. CVPR (2025)
20. Helber, P., Bischke, B., Dengel, A., Borth, D.: EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification. JSTAEORS (2019)
21. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. In: Proc. NeurIPS Deep Learning Workshop (2014)
22. Hu, Z., Wei, Y., Shen, L., Wang, Z., Li, L., Yuan, C., Tao, D.: Sparse model inversion: Efficient inversion of vision transformers for data-free applications. In: Proc. ICML (2024)
23. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: Proc. ICML (2015)
24. Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Zitnick, C.L., Girshick, R.: Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In: Proc. CVPR (2017)

25. Kataoka, H., Hayamizu, R., Yamada, R., Nakashima, K., Takashima, S., Zhang, X., Martinez-Noriega, E.J., Inoue, N., Yokota, R.: Replacing labeled real-image datasets with auto-generated contours. In: Proc. CVPR (2022)
26. Kataoka, H., Matsumoto, A., Yamagata, E., Yamada, R., Inoue, N., Satoh, Y.: Formula-driven supervised learning with recursive tiling patterns. In: Proc. ICCV Workshop on Human-centric Trustworthy Computer Vision (2021)
27. Kataoka, H., Okayasu, K., Matsumoto, A., Yamagata, E., Yamada, R., Inoue, N., Nakamura, A., Satoh, Y.: Pre-training without natural images. In: Proc. ACCV (2020)
28. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Proc. ICLR (2015)
29. Krause, J., Deng, J., Stark, M., Li, F.F.: Collecting a large-scale dataset of fine-grained cars. In: Proc. CVPR-W (2013)
30. Li, Z., Chen, M., Xiao, J., Gu, Q.: Psaq-vit v2: Toward accurate and general data-free quantization for vision transformers. IEEE TNNLS (2024)
31. Li, Z., Ma, L., Chen, M., Xiao, J., Gu, Q.: Patch similarity aware data-free quantization for vision transformers. In: Proc. ECCV (2022)
32. Liu, X., Zhou, J., Kong, T., Lin, X., Ji, R.: Exploring target representations for masked autoencoders. In: Proc. ICLR (2024)
33. Lopes, R.G., Fenu, S., Starner, T.: Data-free knowledge distillation for deep neural networks. arXiv:1710.07535 (2017)
34. Maji, S., Rahtu, E., Kannala, J., Blaschko, M., Vedaldi, A.: Fine-grained visual classification of aircraft. arXiv:1306.5151 (2013)
35. Micaelli, P., Storkey, A.J.: Zero-shot knowledge transfer via adversarial belief matching. In: Proc. NeurIPS (2019)
36. Nakamura, R., Kataoka, H., Takashima, S., Noriega, E.J.M., Yokota, R., Inoue, N.: Pre-training vision transformers with very limited synthesized images. In: Proc. ICCV (2023)
37. Nakamura, R., Tadokoro, R., Yamada, R., Asano, Y.M., Laina, I., Rupprecht, C., Inoue, N., Yokota, R., Kataoka, H.: Scaling backwards: Minimal synthetic pre-training? In: Proc. ECCV (2024)
38. Nakashima, K., Kataoka, H., Matsumoto, A., Iwata, K., Inoue, N., Satoh, Y.: Can vision transformers learn without natural images? In: Proc. AAAI (2022)
39. Nayak, G.K., Mopuri, K.R., Shaj, V., Babu, R.V., Chakraborty, A.: Zero-shot knowledge distillation in deep networks. In: Proc. ICML (2019)
40. Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: Proc. ICVGIP (2008)
41. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Howes, R., Huang, P.Y., Xu, H., Sharma, V., Li, S.W., Galuba, W., Rabbat, M., Assran, M., Ballas, N., Synnaeve, G., Misra, I., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning robust visual features without supervision. TMLR (2024)
42. Parkhi, O.M., Vedaldi, A., Zisserman, A., Jawahar, C.V.: Cats and dogs. In: Proc. CVPR (2012)
43. Patel, G., Mopuri, K.R., Qiu, Q.: Learning to retain while acquiring: Combating distribution-shift in adversarial data-free knowledge distillation. In: Proc. CVPR (2023)
44. Raikwar, P., Mishra, D.: Discovering and overcoming limitations of noise-engineered data-free knowledge distillation. In: Proc. NeurIPS (2022)
45. Ramachandran, A., Kundu, S., Krishna, T.: Clamp-vit: Contrastive data-free learning for adaptive post-training quantization of vits. In: Proc. ECCV (2024)

46. Ranzinger, M., Heinrich, G., Kautz, J., Molchanov, P.: Am-radio: Agglomerative vision foundation model – reduce all domains into one. In: Proc. CVPR (2024)
47. Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y.: Fitnets: Hints for thin deep nets. In: Proc. ICLR (2015)
48. Roth, K., Thede, L., Koepke, A.S., Vinyals, O., Henaff, O.J., Akata, Z.: Fantastic gains and where to find them: On the existence and prospect of general knowledge transfer between any pretrained model. In: Proc. ICLR (2024)
49. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. IJCV (2015)
50. Sarıyıldız, M.B., Alahari, K., Larlus, D., Kalantidis, Y.: Fake it till you make it: Learning transferable representations from synthetic imagenet clones. In: Proc. CVPR (2023)
51. Sariyildiz, M.B., Kalantidis, Y., Larlus, D., Alahari, K.: Concept generalization in visual representation learning. In: Proc. ICCV (2021)
52. Sarıyıldız, M.B., Weinzaepfel, P., Lucas, T., de Jorge, P., Larlus, D., Kalantidis, Y.: DUNE: distilling a universal encoder from heterogeneous 2d and 3d teachers. In: Proc. CVPR (2025)
53. Sariyildiz, M.B., Weinzaepfel, P., Lucas, T., Larlus, D., Kalantidis, Y.: UNIC: universal classification models via multi-teacher distillation. In: Proc. ECCV (2024)
54. Shi, B., Zhang, X., Wang, Y., Li, J., Dai, W., Zou, J., Xiong, H., Tian, Q.: Hybrid distillation: Connecting masked autoencoders with contrastive learners. In: Proc. ICLR (2024)
55. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: Proc. ECCV (2012)
56. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., Massa, F., Haziza, D., Wehrstedt, L., Wang, J., Darcet, T., Moutakanni, T., Sentana, L., Roberts, C., Vedaldi, A., Tolan, J., Brandt, J., Couprie, C., Mairal, J., Jégou, H., Labatut, P., Bojanowski, P.: DINOv3. arXiv:2508.10104 (2025)
57. Takashima, S., Hayamizu, R., Inoue, N., Kataoka, H., Yokota, R.: Visual atoms: Pre-training vision transformers with sinusoidal waves. In: Proc. CVPR (2023)
58. Tang, H., Jia, K.: A new benchmark: On the utility of synthetic data with blender for bare supervised learning and downstream domain adaptation. In: Proc. CVPR (2023)
59. Tian, Y., Krishnan, D., Isola, P.: Contrastive representation distillation. In: Proc. ICLR (2020)
60. Touvron, H., Cord, M., Jegou, H.: DeiT III: Revenge of the ViT. In: Proc. ECCV (2022)
61. Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., Hénaff, O., Harmsen, J., Steiner, A., Zhai, X.: Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv:2502.14786 (2025)
62. Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., Belongie, S.: The iNaturalist species classification and detection dataset. In: Proc. CVPR (2018)
63. Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: SUN database: Large-scale scene recognition from abbey to zoo. In: Proc. CVPR (2010)
64. Yamada, R., Hara, K., Kataoka, H., Makihara, K., Inoue, N., Yokota, R., Satoh, Y.: Formula-supervised visual-geometric pre-training. In: Proc. ECCV (2024)

65. Yamada, R., Takahashi, R., Suzuki, R., Nakamura, A., Yoshiyasu, Y., Sagawa, R., Kataoka, H.: Mv-fractaldb: Formula-driven supervised learning for multi-view image recognition. In: Proc. IROS (2021)
66. Ye, J., Ji, Y., Wang, X., Gao, X., Song, M.: Data-free knowledge amalgamation via group-stack dual-gan. In: Proc. CVPR (2020)
67. Yin, H., Mallya, A., Vahdat, A., Alvarez, J.M., Kautz, J., Molchanov, P.: See through gradients: Image batch recovery via gradinversion. In: Proc. CVPR (2021)
68. Yin, H., Molchanov, P., Alvarez, J.M., Li, Z., Mallya, A., Hoiem, D., Jha, N.K., Kautz, J.: Dreaming to distill: Data-free knowledge transfer via deepinversion. In: Proc. CVPR (2020)
69. Ypsilantis, N.A., Chen, K., Araujo, A., Chum, O.: UDON: Universal dynamic online distillation for generic image representations. In: Proc. NeurIPS (2024)
70. Zagoruyko, S., Komodakis, N.: Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In: Proc. ICLR (2017)
71. Zbontar, J., Jing, L., Misra, I., LeCun, Y., Deny, S.: Barlow twins: Self-supervised learning via redundancy reduction. In: Proc. ICML (2021)
72. Zhai, X., Puigcerver, J., Kolesnikov, A., Ruyssen, P., Riquelme, C., Lucic, M., Djolonga, J., Pinto, A.S., Neumann, M., Dosovitskiy, A., Beyer, L., Bachem, O., Tschannen, M., Michalski, M., Bousquet, O., Gelly, S., Houlsby, N.: A large-scale study of representation learning with the visual task adaptation benchmark. arXiv:1910.04867 (2020)
73. Zhang, W., Liu, Y., Ran, W., Ma, C.: Cross-architecture distillation made simple with redundancy suppression. In: Proc. ICCV (2025)
74. Zhang, X., Qin, H., Ding, Y., Gong, R., Yan, Q., Tao, R., Li, Y., Yu, F., Liu, X.: Diversifying sample generation for accurate data-free quantization. In: Proc. CVPR (2021)
75. Zhao, K., Zhuang, Z., Zhang, M., Guo, C., Shu, Y., Yang, B.: Enhancing diversity for data-free quantization. In: Proc. CVPR (2025)
76. Zhong, Y., Zhou, Y., Zhang, Y., Sui, W., Li, S., Li, Y., Chao, F., Ji, R.: Semantic alignment and reinforcement for data-free quantization of vision transformers. In: Proc. ICCV (2025)
77. Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., Torralba, A.: Semantic understanding of scenes through the ADE20k dataset. IJCV (2019)
78. Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., Kong, T.: iBOT: Image BERT pre-training with online tokenizer. In: Proc. ICLR (2022)