DYNA-DINO: EFFICIENT VIT DISTILLATION VIA ADAPTIVE REPRESENTATION ANCHORING

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision Foundation Models (VFMs) with Vision Transformer (ViT) backbones, such as DINOv2, have become essential for downstream tasks like object recognition and semantic segmentation. The immense computational requirements of backbones often necessitate distillation into smaller architectures for edge deployment. Feature-based knowledge distillation (KD) often suffers from the teacher-student gap; the student struggles to imitate teacher's complex feature map due to its limited capacity. To mitigate this bottleneck, we propose **Dyna-DINO**: Efficient ViT Distillation Via Adaptive Representation Anchoring, a training curriculum for ViT feature-based knowledge distillation. By utilizing the teacher's intermediate feature maps as a sequence of progressively more difficult targets, our curriculum allows the student to build a foundational representation before tackling higher-level abstractions. Our results demonstrate that this paradigm significantly accelerates convergence through adaptive difficulty selection across various student model sizes and dataset scales. With our curriculum, the Dyna-DINO distilled ViT-S achieves 90.1% accuracy on ImageNet-100, a $+12.24\%$ improvement compared with baseline. On ImageNet-1K, Dyna-DINO achieves $+3.9\%$ and $+6.09\%$ improvement for the instance retrieval task on the Oxford and Paris datasets, $+1.93\%$ improvements on semantic segmentation task, as well as meaningful performance gain on classification task. Furthermore, the curriculum enables 25.1% savings in training FLOPs and 21% savings in training time on ImageNet-100 by implementing early-stopping for teacher inference during the initial stages of training.

1 INTRODUCTION

Vision Transformers (ViT) have transformed computer vision by replacing convolutional hierarchies with spatial self-attention mechanisms that treat image patches as tokens (Dosovitskiy et al., 2021). While large-scale Vision Foundation Models (VFMs) like DINOv2 achieve state-of-the-art performance, their massive parameter counts—often reaching the scale of ViT-Giant (2B parameters) or ViT-Huge (700M parameters)—make them impractical for deployment on resource-constrained edge devices. Knowledge distillation (KD) is a common solution to distill these "teacher" models into compact "student" models, such as ViT-Small (22m parameters) (Touvron et al., 2021). Feature-based KD is especially effective for ViTs, as it forces the student to mimic the teacher's intermediate latent representations rather than just final classification logits (Romero et al., 2014). This approach allows the distilled student to retain the teacher's versatility across various downstream tasks, including classification, retrieval, and segmentation.

However, a fundamental challenge remains: the teacher-student gap. Research suggests that as the teacher model grows larger, the performance of a small student model often degrades (Guo et al., 2020). Because the student has significantly lower representational capacity and a lower-rank feature space than the teacher. Attempting to match the complex, high-dimensional feature maps of a 40-layer ViT-G teacher in a single step causes unstable training and slow convergence; the student struggles to learn the teacher's final abstractions before it has mastered basic spatial structures. To bridge this gap, existing literature has explored several strategies to "soften" the distillation target. A primary approach involves intermediate layer matching, where the student is provided with "hints" from the teacher's intermediate layers rather than just the final representation (Hinton et al., 2015). For example, Patient Knowledge Distillation (PKD) introduced fixed-layer selection strate-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

gies—such as matching every k-th layer (PKD-Skip) or the final k layers (PKD-Last)—to ensure the student captures the hierarchical transformation of information (Sun et al., 2019) throughout the network. Other frameworks, such as ViTKD (Yang et al., 2022), use projection heads—often linear or MLP-based layers—to map the student's intermediate lower-dimensional features into the teacher's intermediate high-dimensional manifold to perform feature map generation and mimicking, attempting to resolve the physical dimension mismatch and apply hidden supervision. Relational distillation attempts to mitigate this by matching the similarity between patches rather than their raw values (Park et al., 2019) (Yan et al., 2025). Despite these advancements, most existing methods still rely on a static mapping schedule, where the student is forced to align with complex deep-layer features from the very beginning of training, or mapping the student and teacher's intermediate features manually. This "all-at-once" approach fails to account for the student's evolving capacity, and in the scenario where teacher and student have a large size mismatch the layer-mapping becomes arbitrary.

To mitigate this, we draw inspiration from curriculum learning: a training strategy that introduces concepts in an "easy-to-hard" progression (Bengio et al., 2009). The similarity analysis reveals that the shallower layers of ViT teacher produce feature maps with a higher similarity score with the final student feature map on earlier training epochs, and the similarity peak gradually shifts to the final teacher feature map as training proceeds. According to this finding, we hypothesize that similar feature maps are easier to learn for the student, thus narrowing the teacher-student gap. By treating the teacher's shallower, more reconstructive layers as early and accessible targets, and gradually sweeping toward deeper, semantic layers, we guide the student through a structured learning path that accelerates convergence and improves final feature alignment. With this assumption, we propose **Dyna-DINO**: a training curriculum that gradually switches the training target across intermediate teacher features, with early stopping (controlled by a similarity threshold) pacing the curriculum With experiments on both ImageNet-100 and ImageNet-1K, our distillation curriculum achieves remarkable convergence speed-ups and substantial savings in training FLOPs and wall-time. Evaluations on semantic segmentation, image retrieval and image classification demonstrate that the distilled model retains the ability to adapt to a variety of downstream tasks. In summary, our contributions are as follows:

1. We propose **Dyna-DINO**: a layer-skipping curriculum for feature-based distillation for vision transformers, without the need for manual assignment or intermediate feature selection.

2. A thorough analysis on the robustness for the **Dyna-DINO** curriculum as well as the effect of using intermediate feature maps to supervise student model's features with knowledge distillation. We furthermore verify and evaluate the distilled model on image, instance, and pixel level tasks.

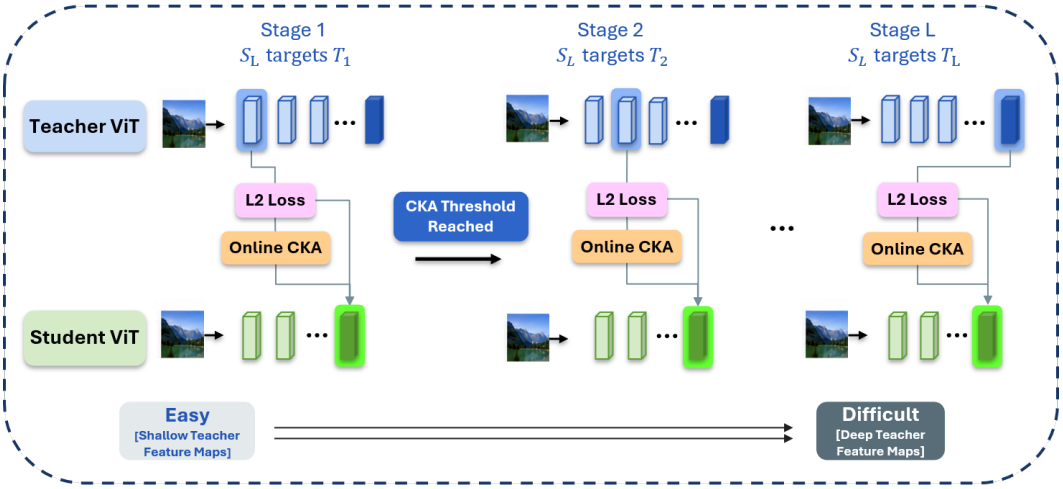

Figure 1: Overview of Dyna. Rather than supervising the student against a fixed teacher block from the start, our curriculum advances the supervisory target through the teacher's feature maps shallow-to-deep based on online CKA alignment, building student representations progressively.

2 Related Work

Vision Transformer Vision Transformers (ViTs) (Dosovitskiy et al., 2021) have achieved state-of-the-art performance in a wide array of downstream applications. Due to their exceptional scalability with respect to dataset size and model capacity, ViTs have become the preferred backbone for modern Vision Foundation Models (VFMs), such as DINOv2 (Oquab et al., 2023) and CLIP (Radford et al., 2021). A key characteristic of the ViT architecture is in its structural hierarchy: shallower blocks tend to capture local spatial details, whereas deeper layers specialize in extracting dense semantic abstractions (Yang et al., 2022; Raghu et al., 2021). When trained in a self-supervised manner, these models exhibit generalization capabilities that enable high performance in zero-shot or frozen-backbone scenarios, including image classification, semantic segmentation, and instance retrieval, without requiring extensive task-specific fine-tuning.

Knowledge Distillation While high-capacity ViT backbones offer state-of-the-art performance, their computational requirements necessitate distillation for resource-constrained environments. Knowledge Distillation (KD) (Hinton et al., 2015) (Gou et al., 2021)bridges this gap by transferring knowledge from large-scale teachers to efficient students. Unlike logit-based methods (Sun et al., 2024) that focus on task-specific output distributions, we utilize feature-based KD (Yang et al., 2022) (Yang et al., 2023) (Chen et al., 2021), which minimizes the MSE between student and teacher feature maps via a linear projection layer. This paradigm is more robust and task-agnostic, making it ideal for distilling foundation models where final projection heads may be unavailable. By focusing on intermediate representations, our method ensures the distilled student retains a rich, general-purpose feature space capable of supporting diverse downstream tasks, from object recognition to dense prediction.

Teacher-Student Gap Due to the mismatch in capacity between the teacher and student models, a larger teacher model does not necessarily lead to a stronger distilled student, and, in some cases, can hamper the student model's performance. This phenomenon is referred to as the teacher-student gap (Tian et al., 2026). Several prior works propose solutions to reduce the teacher's ability to be more compatible with student's capacity (Jin et al., 2019) , or to address the capacity mismatch by examining the gradient similarity (Zhu & Wang, 2021). Normalization for the logits has also been shown to be effective for the logit-based distillation (Cho & Hariharan, 2019). Other works indicate that the usage of additional middle-sized teacher models (Mirzadeh et al., 2020) or intermediate teacher checkpoints during pretraining is also useful for reducing the gap and helpful for distillation into the student (Li et al., 2022). In summary, the approaches from prior works mainly involve processing the output from the teacher model in ways that are more compatible with the student's capacity.

Curriculum Learning Curriculum Learning (CL) (Bengio et al., 2009) improves training by presenting samples in increasing order of difficulty. While CL has recently been used to boost the efficiency of Vision Foundation Models (Zhang et al., 2025; Lu et al., 2025), its application to knowledge distillation remains limited. Existing CL-KD strategies, such as logit temperature scaling (Li et al., 2022) or progressive layer matching (Wang et al., 2018), often struggle with feature-based distillation or rely on arbitrary block-wise assignments in situations when teacher and student architectures differ significantly in depth. To address these limitations, we design a curriculum that leverages the teacher's full range of intermediate features. By ordering these features according to an adaptive similarity metric, our approach eliminates the need for manual layer alignment, facilitating a more natural and efficient transfer of knowledge across disparate architectures

3 Recipe for Efficient ViTKD

3.1 Shifted Similarity

For our baseline, we adopt a standard feature-based KD paradigm that minimizes the L_2 distance between the final feature maps of the student and teacher architectures. In this setup, the teacher's final representation remains the static target throughout the entire training duration. Given that the ViT-G teacher comprises of 40 blocks, whereas the ViT-S student has only 12, a fundamental question arises:

Is the final teacher feature map the most accessible target for the student to learn?

To investigate this, we perform a layer-wise similarity analysis between the student's final feature map and all intermediate feature maps of the teacher during baseline training (see Figure 2). We utilize Centered Kernel Alignment (CKA) as our metric, calculating scores directly on the raw feature maps without intermediate linear projections.

The resulting similarity landscape reveals that the final teacher layer is not always the most similar to the student's output, despite being the sole training target. Instead, we observe a distinct temporal shift: shallower teacher layers exhibit significantly higher similarity during the initial stages of training, with the peak of this similarity distribution gradually advancing toward the final teacher block as optimization progresses. Operating on the premise that *feature similarity correlates with learning ease*, these findings suggest that a static target is sub-optimal. Rather than focusing on a single layer's representation, the student should adaptively navigate the full range of the teacher's representational space, dynamically selecting the "easiest" target at each stage of the distillation process.

3.2 The Layer-Skipping Curriculum

The observed shift in similarity patterns provides the core motivation for our proposed training curriculum. We redefine the complexity of the distillation task by mapping it to the teacher's depth: shallower teacher feature maps serve as initial, more accessible learning targets, while deeper maps represent increasingly complex semantic abstractions. The training process begins with the student imitating the teacher's first ViT block. Rather than following a fixed, manual schedule, we introduce an adaptive progression mechanism controlled by a CKA similarity threshold (τ). Throughout the distillation process, we monitor the "online" CKA similarity between the student's terminal feature map and the current teacher target. The curriculum advances to the next teacher block if and only if the similarity score reaches the predefined threshold τ. This similarity-driven transition ensures that the student has sufficiently learned the current level of abstraction before attempting to bridge the next gap in the representational hierarchy. This adaptive logic is formally detailed in Algorithm 1.

4 Experiments

4.1 Dataset and Training Setup

Dataset We evaluate the proposed framework on ImageNet-100 (Vinyals et al., 2016) and the large-scale ImageNet-1K (Russakovsky et al., 2015) datasets. To facilitate rapid experiment itera-

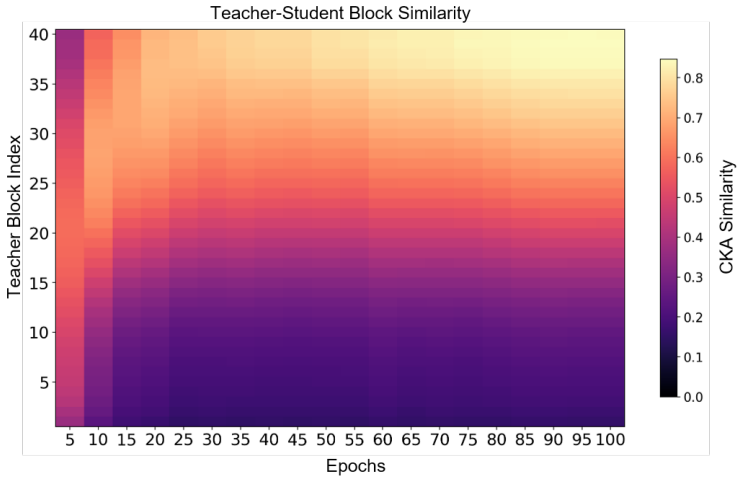

Figure 2: CKA heatmap between the student model's last feature map and all of the teacher's intermediate feature maps during training. Student checkpoints are saved every 5 epochs, and the CKA score is calculated across a subset of validation dataset.

tion, we utilize ImageNet-100 for all ablation studies and parameter tuning. To assess the quality and transferability of the distilled representations, we evaluate performance across several downstream tasks: semantic segmentation is measured on the ADE20K dataset (Zhou et al., 2017; 2019), while instance retrieval is benchmarked on the revisited Oxford and Paris datasets (Radenović et al., 2018).

Training Setup For all distillation experiments, we adopt the standard data augmentation pipeline from (Lightly Team, 2025). Optimization is performed using the LARS optimizer (You et al., 2017) with a global batch size of 256. The base learning rate is set to 9.0 for ImageNet-100 experiments and 6.0 for ImageNet-1K. Our teacher model utilizes a ViT-G backbone pre-trained with the DINOv2 objective and registers (Oquab et al., 2023). For the student models, we evaluate both ViT-Small (ViT-S) and ViT-Tiny (ViT-T), both of which are randomly initialized without registers. Training and evaluation are conducted across a heterogeneous compute cluster featuring NVIDIA L40S, Nvidia RTX A6000, NVIDIA L40, and GeForce RTX 2080 Ti GPUs. All experiments trained on ImageNet-100 cost 12-14 NVIDIA L40S hours, and all ImageNet-1K distillations take 300-400 NVIDIA L40S hours. For all the experiments result tables, we refer "Fixed-anchor" as the baseline for consistently anchoring the student with teacher feature map, "Dyna" as our Dyna-DINO method, and "ViTKD" as another baseline from Yang et al. (2022).

Algorithm 1 CKA-Triggered Progressive Distillation Curriculum

Require: Teacher features $\{T_1, \ldots, T_M\}$, student S, threshold τ, patience E_{max}

$m \leftarrow 1, \quad e \leftarrow 0$ {current target block, epochs on it}
for epoch $= 1$ **to** N_{epochs} **do**
  Sample batch B
  $score \leftarrow \mathrm{CKA}\big(S_{last}(B),\, T_m(B)\big)$
  if $m < M$ **and** ($score \geq \tau$ **or** $e \geq E_{max}$) **then**
    $m \leftarrow m + 1, \quad e \leftarrow 0$ {advance curriculum}
  else
    $e \leftarrow e + 1$
  end if
  $\mathcal{L} \leftarrow \mathrm{MSE}\big(S_{last}(B),\, T_m(B)\big)$
  Update S via $\nabla\mathcal{L}$
end for
return S =0

4.2 The Layer-Skipping Curriculum Saves Training Time and FLOPs

We first validate the effectiveness of our curriculum on ImageNet-100. The distillation objective is defined as:

$$\mathcal{L}_{\text{distill}} = \mathrm{MSE}\left(P(S_{\text{feat}}), T_{\text{feat}}\right) + 0.05 \cdot \mathrm{MSE}\left(P(S_{\text{cls}}), T_{\text{cls}}\right) \tag{1}$$

Following (Zhou et al., 2022), we utilize a single-layer linear projector P to align the student's hidden dimensions with those of the teacher; this projector is shared between the patch tokens and the CLS token. We maintain a constant weight of 0.05 for the CLS token loss throughout all experiments to ensure a consistent comparison.

We evaluate convergence speed by performing linear probing on student checkpoints every five epochs. As illustrated in Figure 3 (left), Dyna distillation demonstrates significantly faster convergence than the baseline from the earliest stages of training. This acceleration suggests that shallower teacher feature maps effectively serve as "easier" learning targets, allowing the student to establish a stable representational foundation before tackling more complex objectives.

The progression of the curriculum, visualized in Figure 3 (right), reveals a distinct temporal pattern. The student transitions rapidly through initial teacher blocks which primarily encode local spatial information, while dwelling significantly longer in the deeper layers. This behavior suggests that deeper layers contain denser semantic knowledge that needs more steps to learn. This evidence validates our assumption of depth serving as a reliable proxy for target difficulty.

To evaluate the generalizability of our curriculum, we investigate its performance across different student capacities: ViT-Tiny (12 blocks, ∼6M parameters) and ViT-Small (12 blocoks, ∼22M pa-

Table 1: Performance and efficiency comparison between baseline distillation and Dyna. Students (ViT-S, ViT-Tiny) are distilled from a ViT-G teacher on ImageNet-100 for 100 epochs. Dyna utilizes a CKA-threshold of 0.85. Efficiency metrics (FLOPs and Train Time) represent the reduction in teacher computational overhead during the distillation process.

Experiments	**Lin. Probe** ↑ (patch token)	**Lin. Probe** ↑ (CLS token)	**FLOPs Saving**	**Train Time Saving**
Fixed-anchor (ViT-S)	77.86%	77.12%	—	—
ViTKD (ViT-S) (Yang et al., 2022)	70.74%	66.88%	—	—
Fixed-anchor (ViT-Tiny)	75.90%	76.04%	—	—
Dyna (ViT-S)	**90.10**%	**89.66**%	25.1%	21%
Dyna (ViT-Tiny)	**81.76**%	**82.02**%	28.82%	22.5%

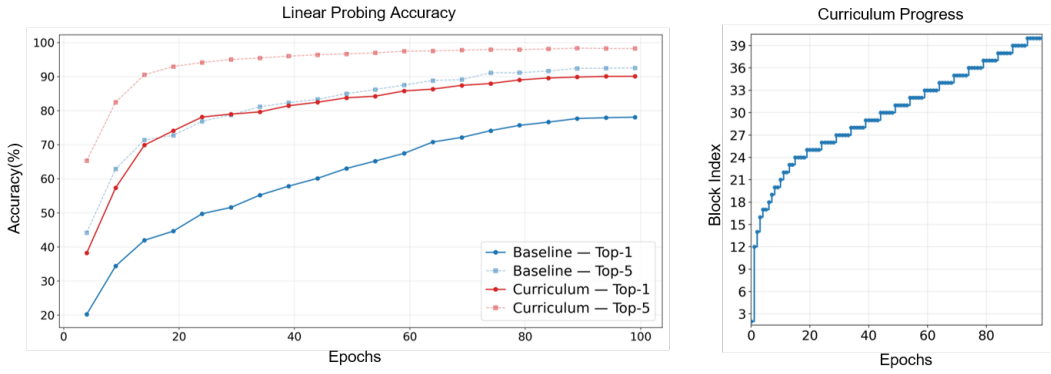

Figure 3: Left: the linear probing accuracy convergence comparison for baseline and Dyna. Right: the curriculum progress visualization. The student checkpoint for linear probing evaluation is saved for every 5 epochs. "Baseline" refers to Fixed-anchor and "Curriculum" is our Dyna method.

rameters). As detailed in Table 1, Dyna consistently enhances linear probing accuracy for both architectures, with ViT-S reaching 90.10% and ViT-Tiny achieving 81.76%. Critically, these performance gains are accompanied by significant computational savings; the adaptive early-stopping for teacher inference achieves up to a 28.8% reduction in training FLOPs and a 22.5% reduction in total training time.

4.3 Evaluating Representational Quality via Downstream Tasks

While the accelerated convergence of the proposed curriculum is demonstrated through image-level linear probing, we further evaluate the distilled models on diverse downstream tasks to assess the generalizability of the learned representations. We conduct evaluations on semantic segmentation and instance retrieval to verify that the student model preserves both pixel-level details and fine-grained instance features. These experiments confirm that our curriculum-based distillation not only speeds up training but also produces a versatile backbone capable of adapting to a wide range of applications.

To evaluate instance-level representations, we perform image retrieval experiments on the Oxford and Paris datasets. For each image, the model backbone extracts a global embedding; we then compute the similarity scores between the query image embedding and the database embeddings. These results are ranked by similarity to calculate the mean Average Precision (mAP). As shown in Table 2, our Dyna-distilled ViT-S and ViT-Tiny outperform the baseline by a significant margin across all three difficulty levels. These results demonstrate that the proposed curriculum effectively preserves the fine-grained structural details required for precise instance matching.

Table 2: Instance recognition performance comparison (mAP) on Oxford and Paris datasets. The "Mean" column represents the average mAP across Easy, Medium, and Hard difficulty levels.

	roxford5k				rparis6k			
Training method	E	M	H	**Mean**	E	M	H	**Mean**
Fixed-anchor (ViT-S)	10.28	8.80	2.15	7.08	24.83	21.29	6.63	17.79
ViTKD (ViT-S)	7.48	6.82	1.37	5.22	28.55	21.20	7.25	18.79
Fixed-anchor (ViT-Tiny)	8.04	7.72	1.84	5.87	21.35	19.42	6.59	15.79
Dyna (ViT-S)	**22.56**	**17.30**	**4.81**	**14.89**	**53.84**	**40.99**	**15.97**	**36.93**
Dyna (ViT-Tiny)	**14.70**	**12.15**	**2.84**	**9.90**	**43.39**	**33.36**	**12.38**	**29.71**

We evaluate performance on ADE20K using three protocols: Linear Segmentation, Encoder-Only Mask Transformer (EOMT), and Multi-Scale (MS) inference. The results are summarized in Table 3.

In the Linear Segmentation setting, which measures the linear separability of frozen features, Dyna improves the ViT-S mIoU from 12.15% to 20.53%. When utilizing the EOMT head to evaluate how features interact within a Transformer-native decoder, the performance gap remains significant, with the Dyna-distilled ViT-S reaching 38.10% mIoU compared to the baseline's 24.49%. Finally, under Multi-Scale (MS) evaluation (a protocol that tests robustness to scale variations) the Dyna-distilled ViT-S achieves 39.36% mIoU, representing a 14.74% absolute improvement. These consistent gains across different decoding architectures and inference scales suggest that navigating the teacher's representational hierarchy during training allows the student to retain denser spatial and semantic information than standard distillation regimes.

Table 3: Semantic segmentation performance comparison for distillation on ImageNet-100. Results are reported in mean Intersection over Union (mIoU) for linear segmentation, EOMT validation, and multi-scale (MS) evaluation.

Method	**Linear Seg.** ↑ (mIoU)	**EOMT val** ↑ (mIoU)	**MS** ↑ (mIoU)
Fixed-anchor (ViT-S)	12.15%	24.49%	24.62%
ViTKD (ViT-S) (Yang et al., 2022)	11.51%	14.94%	14.75%
Fixed-anchor (ViT-Tiny)	9.51%	15.41%	14.88%
Dyna (ViT-S)	**20.53**%	**38.10**%	**39.36**%
Dyna (ViT-Tiny)	**14.71%**	**21.67**%	**21.82%**

4.4 Effectiveness on Large Dataset

Following the demonstrated effectiveness on ImageNet-100, we evaluate the scalability of the Dyna curriculum on ImageNet-1K. We introduce a more challenging baseline; in addition to only anchoring the student's final feature map (which we refer to as Fixed-anchor(1 Anchor), we anchor the student's 1st, 6th, and the final feature map and name this baseline Fixed-anchor (3 Anchors).

The advantages of Dyna remain evident in tasks requiring high-fidelity features. As shown in Table 4, Dyna achieves remarkable performance across the Oxford and Paris retrieval benchmarks. Specifically, the Dyna-distilled ViT-S achieves a mean mAP improvement of 3.9% on the Oxford dataset and 6.09% on the Paris dataset. These results suggest that even at scale, Dyna effectively transfers the structural nuances which are critical for fine-grained instance matching.

In semantic segmentation (Table 5) and linear probing (Table 6), we observe that Dyna produces competitive results with the standard distillation baseline. In semantic segmentation, the Dyna-Distilled ViT-S shows significantly better performance on linear segmentation, EOMT, and Multi-scale evaluation. Similarly, linear probing accuracies on ImageNet-1K demonstrate the effectiveness with ViT-S reaching 78.10% (vs. 77.63% and 72.31% baseline) and ViT-Tiny reaching 64.62% (vs. 64.4% baseline). We have an interesting observation that while fixed-anchor(3 anchors) obtains

Table 4: Results for ImageNet-1K. Performance comparison (mAP) on Oxford and Paris datasets. The "Mean" column represents the average mAP across Easy (E), Medium (M), and Hard (H) difficulty levels.

	roxford5k				rparis6k			
Training method	E	M	H	**Mean**	E	M	H	**Mean**
Fixed-anchor (S, 1An)	27.34	19.36	4.86	17.19	62.70	50.46	27.80	46.99
Fixed-anchor (S, 3An)	33.85	23.31	4.64	20.6	58.62	47.25	24.35	43.41
ViTKD (ViT-S)	3.87	3.88	0.87	2.87	13.85	13.07	5.32	10.75
Fixed-anchor (T, 1An)	19.41	14.16	**2.42**	11.9	50.18	38.47	14.98	34.54
Dyna (S, 3An)	**34.12**	**24.31**	**4.86**	**21.1**	**68.04**	**55.78**	**33.54**	**52.45**
Dyna (T, 3An)	**21.93**	**14.99**	2.40	**13.11**	**55.91**	**42.97**	**17.50**	**35.94**

Table 5: Semantic segmentation performance comparison. Results are reported in mean Intersection over Union (mIoU) for linear segmentation, EOMT validation, and multi-scale (MS) evaluation.

Method	**Linear Seg.** ↑ (mIoU)	**EOMT val** ↑ (mIoU)	**MS** ↑ (mIoU)
Fixed-anchor (ViT-S, 1 Anchor)	29.85%	46.65%	48.22%
Fixed-anchor (ViT-S, 3 Anchors)	27.41%	36.23%	37.20%
ViTKD (ViT-S) (Yang et al., 2022)	4.28%	6.3%	6.19%
Fixed-anchor (ViT-Tiny)	18.92%	33.93%	33.42%
Dyna (ViT-S, 3 Anchors)	**31.78%**	**48.68%**	**50.05%**
Dyna (ViT-Tiny, 3 Anchors)	**20.23%**	**37.37%**	**37.61%**

better performance than fixed-anchor(1 anchors) on image retrieval task, it performs worse on the segmentation and linear probing, which require high linear separability in the underlying representation. Dyna outperforms both baselines on both classification and retrieval tasks, and we attribute this benefit to the fact that Dyna curriculum effectively distills a more versatile and functional representation for downstream tasks.

While ViTKD is able to show a matching performance on ImageNet-100 experiments, it does not converge well in the ImageNet-1K scale training. In the original paper of ViTKD, all the distillation experiments are conducted with teacher/student with similar parameter size(ViT-B teacher with ViT-S student, for example.) We attribute the result to this design choice: the method is designed for teacher and student with similar capability, and when it comes to more challenging scenario with huge teacher-student gap (ViT-G teacher and ViT-S student), the assumption that teacher and students are structural similar no longer stand, and this method could fail.

While achieving competitive performance on global classification and dense prediction tasks, Dyna provides a measured improvement in training efficiency, yielding a 12.4% reduction in teacher FLOPs and a 7.5% decrease in training time for the ViT-Tiny student. Though these computational savings are modest at the ImageNet-1K scale, they are achieved while enhancing the model's feature integrity. Specifically, the substantial gains observed in the image retrieval benchmarks, where Dyna outperforms the baseline by up to 6.09%, serve as a strong proxy for overall representation quality. The results confirm that Dyna produces a feature space that better preserves the complex representation of the foundation model teacher.

4.5 Comparison with Dense One-to-One Layer Alignment

When teacher and student architectures share an identical depth, a common strategy for maximizing supervision is dense intermediate matching: aligning every student block with its corresponding teacher block, an approach often viewed as the the upper-bound for feature-based distillation.

In this section, we compare Dyna against this dense matching baseline. We conduct this experiment using a ViT-S teacher/ ViT-Tiny student and ViT-B teacher/ViT-S student; since both models consist

Table 6: Results for training on ImageNet-1K. Comparison of Baselines and Dyna methods for ViT-S and ViT-Tiny with ViT-G teacher. for both ViT-Tiny and ViT-S the CKA threshold for curriculum is 0.8. Efficiency metrics (FLOPs and Train Time saving) represent the reduction in teacher computational overhead during the distillation process.

Experiments	**Lin. Probe** ↑ (patch token)	**Lin. Probe** ↑ (CLS token)	**FLOPs Saving**	**Train Time Saving**
Fixed-anchor (ViT-S, 1 Anchor)	76.87%	77.63%	—	—
Fixed-anchor (ViT-S, 3 Anchors)	73.37%	72.31%	—	—
ViTKD (ViT-S) (Yang et al., 2022)	15.28%	12.09%	—	—
Fixed-anchor (ViT-Tiny, 1 Anchor)	63.35%	64.4%	—	—
Dyna (ViT-S, 3 Anchors)	**77.03%**	**78.10%**	**9.44%**	**4.83%**
Dyna (ViT-Tiny, 3 Anchors)	**63.37%**	**64.62%**	**12.4%**	**7.5%**

of 12 layers, they allow for a direct, one-to-one feature map alignment. The results in table 7 reveal the efficiency of Dyna; with only 1 projector, Dyna is merely 0.02% behind the dense layer alignment upper bound, which uses 12 projectors in total. In the case of ViT-B teacher with ViT-S student, Dyna even outperforms the dense one to one alignment by 2.86%. This comparison indicates that a structured, adaptive curriculum can achieve or even surpass dense supervision while maintaining much higher structural simplicity.

Table 7: Comparison of distillation alignment strategies. One-to-one alignment refers to matching each teacher layer with corresponding student layer, while Dyna utilizes an adaptive progression curriculum. The Fixed-anchor method aligns the last teacher and student feature map.

Alignment Strategy	**Linear Probing** ↑	**Projectors (Params)**
Fixed-anchor (S→ T)	81.80%	×1 (0.07M)
One-to-one Alignment (S→ T)	83.38%	×12 (0.89M)
Dyna (S→ T)	83.36%	×1 (0.07M)
Fixed-anchor (B→ S)	72.52%	×1 (0.3M)
One-to-one Alignment (B→ S)	83.82%	×12 (3.55M)
Dyna (B→ S)	86.68%	×1 (0.3M)

5 Conclusion, Limitations, and Future Directions

In this work, we propose a layer-skipping curriculum for efficient feature-based knowledge distillation in Vision Transformers (ViTs). Through an automatic similarity-based curriculum, the student model achieves accelerated convergence and learns high-quality representations for diverse downstream tasks. One limitation of our work is the assumption of a white-box teacher model with accessible intermediate features. While many open-weight Vision Foundation Models (VFMs) exist, our method is restricted in scenarios where only the final teacher feature map is available. Furthermore, the CKA threshold utilized for our ImageNet-1K experiments may represent a sub-optimal hyperparameter. A more exhaustive optimization of this threshold could yield further performance improvements. A potential future direction is to expand this framework to cross-architecture scenarios where the teacher and student have different architectures (e.g., Transformer to CNN). Another key direction involves extending our method to distilling across data modalities for multimodal models, such as audio, language, and more.

AI use statement

In this work, we used generative AI tools for create experiment code for plotting graph and reformat dataset, . We have not used generative AI tools for interpret results, and Generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine

hypotheses, design or provide feedback on research methodology or experiments, support qualitative and thematic data analysis, are not applicable to this work. Additionally, we used generative AI tools for improve the paper's readability. We have reviewed all AI-assisted work. All LLM-generated code was verified and tested for correctness by all authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Reproducibility statement

A zip file with necessary code are uploaded. Also, the training setup section in this paper listed the necessary configurations to reproduce the results.

References

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th International Conference on Machine Learning*, 2009.

Defang Chen, Jian-Ping Mei, Yuan Zhang, Can Wang, Yan Feng, and Chun Chen. Cross-layer distillation with semantic calibration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Jianping Gou, Baosheng Yu, Stephen J. Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 2021.

Jia Guo, Minghao Chen, Yao Hu, Chen Zhu, Xiaofei He, and Deng Cai. Reducing the teacher-student gap via spherical knowledge distillation. In *arXiv preprint arXiv:2010.07485*, 2020.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Xiao Jin, Baoyun Peng, Yichao Wu, Yu Liu, Jiaheng Liu, Ding Liang, Junjie Yan, and Xiaolin Hu. Knowledge distillation via route constrained optimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

Zheng Li, Xiang Li, Lingfeng Yang, Borui Zhao, Renjie Song, Lei Luo, Jun Li, and Jian Yang. Curriculum temperature for knowledge distillation. *arXiv preprint arXiv:2211.16231*, 2022.

Lightly Team. Lightlytrain, 2025. URL `https://github.com/lightly-ai/lightly-train`.

Wenquan Lu, Jiaqi Zhang, Hugues Van Assel, and Randall Balestriero. Ditch the denoiser: Emergence of noise robustness in self-supervised learning from data curriculum. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL `https://openreview.net/forum?id=Pa5pKAeAO7`.

Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

Filip Radenović, Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondřej Chum. Revisiting Oxford and Paris: Large-scale image retrieval benchmarking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. 2021.

Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *arXiv:2108.08810*, 2021.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *arXiv preprint arXiv:1409.0575*, 2015.

Shangquan Sun, Wenqi Ren, Jingzhi Li, Rui Wang, and Xiaochun Cao. Logit standardization in knowledge distillation. *arXiv preprint arXiv:2403.01427*, 2024.

Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. Patient knowledge distillation for BERT model compression. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019.

Huiyuan Tian, Bonan Xu, and Shijian Li. Distillation dynamics: Towards understanding feature-based distillation in vision transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

Hugo Touvron, Matthieu Cord, Douze Matthijs, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, 2021.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems*, 2016.

Hui Wang, Hanbin Zhao, Xi Li, and Xu Tan. Progressive blockwise knowledge distillation for neural network acceleration. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, 2018.

Zhaoyi Yan, Kangjun Liu, and Qixiang Ye. Delving deep into semantic relation distillation. *arXiv preprint arXiv:2503.21269*, 2025.

Chuanguang Yang, Xinqiang Yu, Zhulin An, and Yongjun Xu. Categories of response-based, feature-based, and relation-based knowledge distillation. *arXiv preprint arXiv:2306.10687*, 2023.

Zhendong Yang, Zhe Li, Ailing Zeng, Zexian Li, Chun Yuan, and Yu Li. ViTKD: Feature-based knowledge distillation for vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop*, 2022.

Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks. *arXiv preprint arXiv:1708.03888*, 2017.

Jiaqi Zhang, Juntuo Wang, Zhixin Sun, John Zou, and Randall Balestriero. FastDINOv2: Frequency based curriculum learning improves robustness and training speed. *arXiv preprint arXiv:2507.03779*, 2025.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ADE20K dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ADE20K dataset. *International Journal of Computer Vision*, 127(3):302–321, 2019.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. iBOT: Image BERT pre-training with online tokenizer. In *International Conference on Learning Representations*, 2022.

Yizhen Zhu and Yi Wang. Student customized knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

A Appendix

A.1 The Necessity of Progressive Supervision vs. Single-Layer Targets

A potential critique of intermediate feature distillation is that the performance gains may stem from selecting a specific ”optimal” layer rather than the curriculum itself. If a single intermediate layer contained the most transferable knowledge, static distillation from that ”lucky” layer would outperform a progressive curriculum. To investigate this, we conducted an ablation study using a ViT-S teacher and a ViT-Tiny student. We trained twelve separate student models, each supervised exclusively by a different fixed intermediate teacher layer throughout training. This allows us to isolate the contribution of individual layers and determine whether Dyna's effectiveness is derived from a specific static target or the structured transition between them.

As illustrated in Figure 4, while deeper teacher layers generally results in higher feature quality, Dyna outperforms all individual targets. These results suggest that no single frozen intermediate layer is sufficient to bridge the gap between teacher and student as effectively as a curriculum. Instead, the performance benefit is a product of the structured progression itself, which allows the student to learn a hierarchical foundation that any single-layer target lacks.

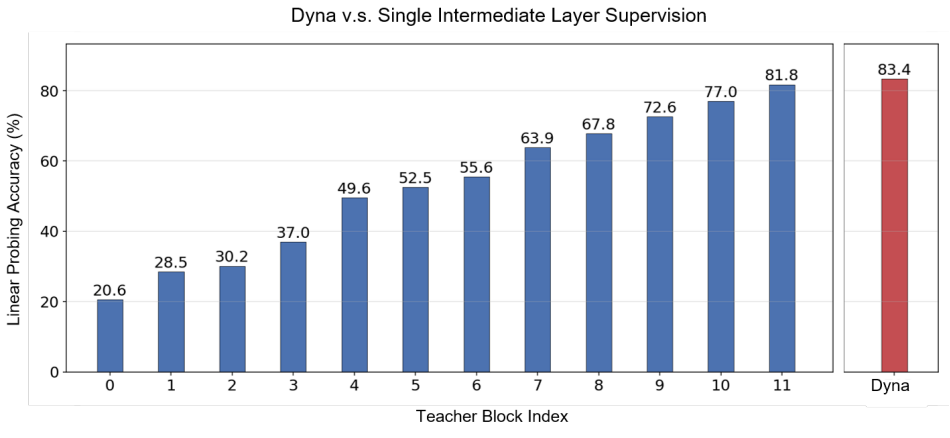

Figure 4: Linear probing accuracy comparison between Dyna and single intermediate layer supervision. Dyna outperforms using any single layer as supervision throughout training, indicating the effectiveness of utilizing a structural curriculum.