Efficient Model Pruning via Selective Layer-wise Distillation with Dynamic CKA-based Weighting

Chi-Dieu Bui[1] and Chien-Quang Le[1, *]
[1]Department of Information Technology, University of Sciences, Hue University, Hue, Vietnam
Email: {22T1020069, lqchien}@husc.edu.vn

***Abstract*—Deploying deep convolutional neural networks in resource-constrained edge environments necessitates aggressive model compression. While iterative block-level pruning paired with multi-stage Knowledge Distillation (KD) is a common strategy, traditional KD approaches rigidly enforce static, uniform loss weightings, leading to representational collapse under severe depth pruning. We propose a novel Multi-stage Knowledge Distillation framework governed by a Dynamic Centered Kernel Alignment (CKA) weighting mechanism. By utilizing CKA to quantify the immediate post-pruning layer-wise representation similarity, our method topology-adaptively modulates distillation intensity, creating a representational "corridor of freedom". This empowers the "student" network to selectively filter crucial knowledge and optimize its feature space autonomously, preventing blind mimicry of the unpruned "teacher". Extensive experiments on closed-set image classification (CIFAR-10/100) and open-set face recognition finetuned on the CASIA-1k/3k and MS1MV2-1k datasets and evaluated on the LFW, CFP-FP, and AgeDB-30 benchmarks demonstrate remarkable scalability. Notably, our framework achieves a near-complete or full performance recovery on the evaluated benchmarks compared to the unpruned teacher, even under an extreme 74% depth reduction (20 blocks pruned). By slashing the recovery phase to just 60 epochs, our approach reduces standard fine-tuning time by 70% (from 200 to 60 epochs). With an Epoch Efficiency Score (S_{eff}) approximately 3.3× higher than traditional baselines, this work establishes a highly effective framework for rapid, high-performance model optimization in resource-constrained Edge AI environments.**

***Index Terms*—Model Compression, Network Pruning, Knowledge Distillation, Centered Kernel Alignment (CKA), Face Recognition**

I. Introduction

Deep convolutional neural networks (CNNs) have driven major advances in image classification [1], [2] and open-set face recognition [3]. However, their massive parameter counts and floating-point operations (FLOPs) hinder deployment in resource-constrained environments, making model compression a critical objective [4].

To address this computational burden, block-level network pruning [5] and multi-stage knowledge distillation [6] can be viewed as complementary strategies for improving model efficiency while maintaining recognition performance. However, a common simplifying design is to assign fixed or uniform weights to distillation losses across stages, which may be suboptimal under aggressive pruning. When a severely compressed student network is required to mimic the high-level representations of a much stronger teacher, the resulting teacher–student capacity gap can hinder effective knowledge transfer and make the teacher feature space difficult for the student to reproduce [7]. This issue is especially concerning in face recognition, where angularly discriminative hyperspherical embeddings are crucial for separating identities [8].

To resolve the limitations of static layer-wise alignment, we propose a paradigm shift in how intermediate features are distilled. By leveraging Centered Kernel Alignment (CKA) [9] as a dynamic measurement of spatial correlation, we introduce a methodology that selectively filters knowledge transfer based on post-pruning representational similarity. In summary, our core contributions are threefold:

- **Dynamic CKA-based Weighting Mechanism:** We propose a novel framework that utilizes intermediate CKA profiles to adaptively assign layer-wise KD weights based on the specific pruned topology. Instead of forced imitation, this mechanism creates a "corridor of freedom" for the pruned student, allowing it to bypass overly complex, high-level teacher features while maintaining structural representation integrity.
- **Accelerated 60-Epoch Training Recipe:** We introduce a highly optimized, phase-scheduled fine-tuning strategy integrating a standard cross-entropy (α-CE) objective with our dynamically weighted multi-stage alignment (β-MSE). Validated strictly on open-set Face Recognition benchmarks (LFW, CFP-FP, AgeDB-30) and general classification datasets (CIFAR-10), this recipe effectively restores True Acceptance Rates (TAR) and baseline accuracy in just 60 epochs—effectively reducing standard training convergence time by 70%.
- **Algorithmic Interpretability via CKA Heatmaps:** We provide an in-depth empirical analysis of our method's behavior through the visualization of CKA heatmaps. By mapping the shifting correlation between student and teacher spaces, we demonstrate that our dynamically guided student learns a more flexible and decoupled representation rather than succumbing to blind memorization.

II. Related Work

Structured depth pruning directly reduces inference cost on general-purpose hardware [10], but removing entire residual

* Corresponding author

979-8-3195-0175-2/26/$31.00 ©2026 IEEE

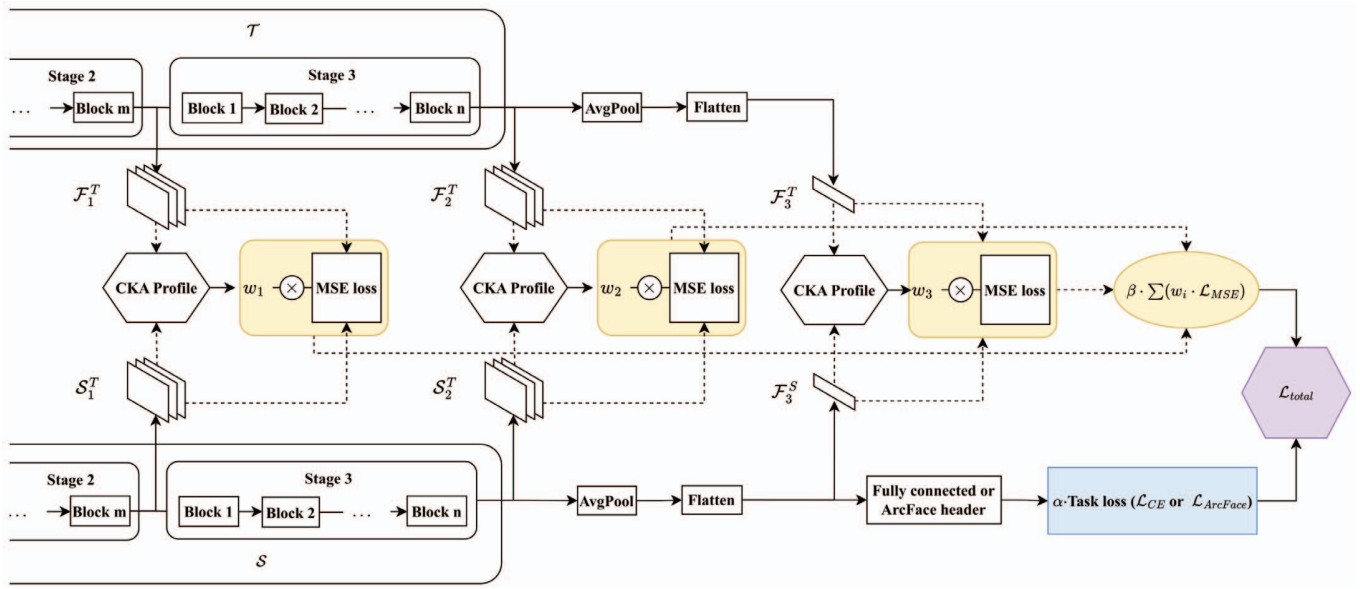

Fig. 1. The proposed dynamic multi-stage knowledge distillation process. The CKA metric dynamically dictates the distillation intensity (w_i), inversely weighting the alignment loss to prioritize the recovery of the most severely degraded representational spaces.

blocks can disrupt hierarchical feature propagation and degrade accuracy, especially in face recognition [11]. Recent advancements in structural pruning [20] and streamlined knowledge distillation [21] have improved generalized compression efficiency. Recent CKA-based pruning methods identify redundant layers more reliably [12]; however, their recovery stage still depends on long fine-tuning schedules, leaving post-pruning optimization computationally expensive.

Knowledge Distillation (KD) [13] and multi-stage feature alignment [6] are commonly used to restore compressed models, yet static layer-wise supervision can over-constrain the student. Prior work shows that rigid teacher imitation may restrict student flexibility [14] and contribute to representation collapse [15], particularly when the capacity gap becomes large. CKA provides a robust measure of representation similarity [9] and has mainly been used for offline analysis [16] or pre-pruning layer selection [12]. In contrast, our method uses CKA online during recovery to dynamically weight distillation, reducing forced mimicry while accelerating convergence.

III. Proposed Method

A. Pruning Setup and Dynamic CKA-Guided Alignment

Let $\mathscr{T}$ and $\mathscr{S}$ denote the highly-parameterized Teacher and the depth-pruned Student networks, respectively (e.g., ResNet-56 for closed-set classification, and an angular-margin optimized iResNet-50 with ArcFace [17] for face recognition). Following the metric-based paradigm [12], we iteratively prune redundant residual blocks exhibiting the lowest Centered Kernel Alignment (CKA) scores, initializing $\mathscr{S}$ with the surviving weights. While this topological reduction satisfies computational constraints, it inherently fractures the forward-pass pathways from the pruning site onwards.

To recover this lost representational capacity without enforcing rigid, uniform feature distillation [18], we introduce a dynamic, forward-conditional alignment framework (Figure 1). Since removing a block alters representations solely at its host and subsequent macro-stages, we explicitly target these impacted boundaries. Immediately following a pruning iteration, we execute a validation forward pass to extract spatial feature maps $F_i^{\mathscr{S}}$ and $F_i^{\mathscr{T}}$ at these stages. Because our block-level pruning strictly reduces depth without altering channel dimensions, $F_i^{\mathscr{S}}$ and $F_i^{\mathscr{T}}$ remain perfectly identical in shape. This inherent structural alignment allows direct CKA and MSE computation by simply flattening the spatial features, eliminating the need for any intermediate projection, pooling, or normalization layers. For the i-th impacted stage, the spatial correlation is quantified via the Linear CKA metric:

$$\mathrm{CKA}_i = \frac{\left\| (F_i^{\mathscr{T}})^T F_i^{\mathscr{S}} \right\|_F^2}{\left\| (F_i^{\mathscr{S}})^T F_i^{\mathscr{S}} \right\|_F \left\| (F_i^{\mathscr{T}})^T F_i^{\mathscr{T}} \right\|_F} \tag{1}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. Empirically, the host stage exhibits the lowest CKA score (indicating severe representational deviation), while deeper layers gradually map the flawed input back to a more stable state (higher CKA).

To optimize the recovery trajectory, we assign distillation weights w_i inversely proportional to the CKA scores. For optimization stability, these weights are normalized across the set of impacted stages $\mathscr{V}$:

$$w_i = \frac{1 - \mathrm{CKA}_i}{\sum_{j \in \mathscr{V}} (1 - \mathrm{CKA}_j)} \tag{2}$$

This mathematically guarantees that $\sum w_i = 1$, bounding the loss magnitude. Functionally, it acts as a self-balancing filter: the distillation penalty peaks at the broken pathway to force

structural repair, while automatically scaling down at deeper layers. This relaxation grants the Student a representational "corridor of freedom," preventing the blind memorization of overly complex mappings. Preceding intact stages are assigned $w_i = 0$ to bypass unnecessary computations. Crucially, to maintain computational efficiency during the accelerated 60-epoch recovery, these topology-adaptive weights w_i are computed only once immediately after pruning and are kept fixed throughout the fine-tuning phase. In practice, this one-time CKA profiling requires a single validation forward pass through both $\mathscr{T}$ and $\mathscr{S}$, taking approximately 40.4 s for ResNet-56 (5,000 images) and 25.2 s for iResNet-50 (4,389 images). This negligible overhead is fully amortized across 60 training epochs, leaving the per-iteration cost identical to that of standard static KD.

B. Efficient 60-Epoch Training Recipe

Traditional pruning recovery protocols require resource-heavy schedules (often > 200 epochs) [12]. Leveraging our CKA-guided mechanism, we propose an accelerated 60-epoch phase-scheduled recipe. The total objective integrates a task-specific loss ($\mathscr{L}_{\text{task}}$: Cross-Entropy or Angular Margin Softmax) with our dynamically weighted alignment loss:

$$\mathscr{L}_{\text{total}} = \alpha \mathscr{L}_{\text{task}} + \beta \sum_{i \in \mathscr{V}} w_i \mathscr{L}_{\text{MSE}} \left(F_i^{\mathscr{S}}, F_i^{\mathscr{T}} \right) \tag{3}$$

where task-specific (α) and alignment (β) hyperparameters rebalance the gradient signals. Empirically, we set $\alpha \in [1,2]$ and dynamically schedule $\beta \in [70,150]$ across phases for CIFAR, while using static $\beta = 500$ ($\alpha = 0$) for face recovery.

To maximize efficiency, the 60-epoch budget is explicitly allocated into a 3-phase strategy: a 5-epoch warm-up to prevent divergence, a 50-epoch main phase using Cosine Annealing [19] for robust CKA-guided distillation, and a 5-epoch cool-down where KD constraints are decoupled to fine-tune task-specific decision boundaries. This structured recipe fully recovers baseline performance, slashing standard training time by over two-thirds.

IV. Experimental Results

A. Setup and Datasets

We evaluate on two image classification benchmarks: CIFAR-10 [22] and CIFAR-100 [22], and three face recognition datasets: LFW [23], CFP-FP [24], and AgeDB-30 [25]. For classification, we report Top-1 Accuracy. For face recognition, we extract 512-dimensional embeddings and report verification accuracy on LFW, and the True Accept Rate (TAR) at a False Accept Rate (FAR) of 10^{-4} on the challenging CFP-FP and AgeDB-30 datasets. All proposed experiments strictly adhere to our accelerated 60-epoch Cosine Annealing schedule, utilizing AdamW or SGD momentum optimizers, directly challenging the computationally intensive 200-epoch step-decay baseline established by the original metric-based pruning framework [12].

B. Performance and Scalability Analysis

We evaluate the structural resilience of our method by comparing classification accuracy and True Acceptance Rate (TAR) under varying degrees of network sparsity (e.g., 1 block vs. 9 blocks pruned).

Closed-Set Classification. Table I details the Top-1 Accuracy on the CIFAR-10 dataset. At very marginal pruning levels (e.g., dropping only 1 block), the remaining representational capacity of the Student is sufficiently large that the performance variance between different Knowledge Distillation paradigms is minimal. Therefore, to clearly illustrate the behavioral divergence when computational resources are strictly constrained, our analysis focuses on moderate (5 blocks) to aggressive (9 blocks) pruning regimes.

When pruning is moderate (5 blocks dropped), the structural damage becomes noticeable. Here, the traditional 200-epoch Standard FT baseline achieves 89.11%. Enforcing strict, uniform feature alignment (Static KD) in an accelerated 60-epoch schedule performs comparably, achieving 89.17%. However, our Dynamic CKA-KD begins to demonstrate its superiority by intelligently relaxing the constraints, yielding 89.43% and effectively outperforming both the static distillation and the resource-intensive baseline.

However, under aggressive pruning (dropping 9 blocks via CKA similarity), a critical divergence emerges. While traditional Static KD achieves 89.21%, our Dynamic CKA-KD intelligently bypasses rigid alignment at terminal layers, achieving the highest accuracy of 89.72% within merely 60 epochs. Notably, this accelerated 60-epoch training outperforms the computationally intensive 200-epoch Standard FT baseline (88.69%). This provides a profound empirical insight: when the capacity gap is severe, optimizing pure feature similarity (via forced static KD) restricts the model. By dynamically establishing a "corridor of freedom," our method prevents the severely pruned Student from acting as a flawed replica, instead allowing it to allocate its limited parameter capacity toward optimizing decision boundaries.

A stage-wise sensitivity analysis confirms this selective approach: Stage 3 exhibits the lowest CKA Gap (0.2295) compared with Stage 1 (0.9023) and Stage 2 (0.3802), validating that deeper layers are more structurally resilient and justifying the assignment of $w_i = 0$ to preceding intact stages.

TABLE I
Top-1 Accuracy on CIFAR-10 (ResNet-56).

Method	Pruned Blks	Epochs	Acc (%)
Teacher (Unpruned)	0	200	87.92
Standard FT [12]	5	200	89.11
Static KD	5	60	89.17
Dynamic CKA-KD (Ours)	**5**	**60**	**89.43**
Standard FT [12]	9	200	88.69
Static KD	9	60	89.21
Dynamic CKA-KD (Ours)	**9**	**60**	**89.72**

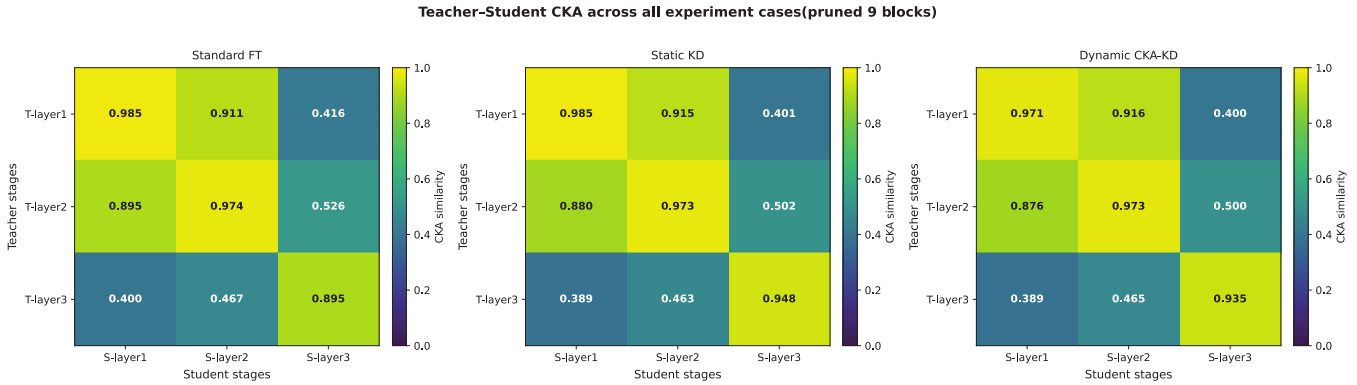

Fig. 2. Cross-network CKA similarity heatmaps between the unpruned Teacher (y-axis) and the pruned Student (x-axis) across different pruning regimes. Left Column: Traditional Static KD forces a strict, constrained diagonal correlation. Right Column: Our Dynamic CKA-KD exhibits a broader, distributed correlation signature, validating the creation of a representational "corridor of freedom" that mitigates structural collapse.

To further assess the proposed weighting strategy as class granularity increases, we extend the evaluation to CIFAR-100 (Table II). While the unpruned Teacher reaches 60.22%, our Dynamic CKA-KD achieves a peak accuracy of 62.06% at the 4-block pruning level (+1.84% improvement), demonstrating a significant "noise filter" effect. At the 5-block milestone, our method maintains 61.81%, matching Static KD and outperforming the 200-epoch baseline. Under aggressive pruning (9 blocks), although Static KD is marginally higher by 0.11%, Dynamic CKA-KD still surpasses the standard fine-tuning baseline by +0.67 percentage points. This indicates that while fine-grained datasets occasionally benefit from stronger teacher anchoring at moderate depths, our accelerated CKA-guided recipe remains robust and highly efficient, significantly reducing computational costs while maintaining superior accuracy over traditional resource-intensive paradigms.

TABLE II
TOP-1 ACCURACY ON CIFAR-100 (RESNET-56).

Method	Pruned Blks	Epochs	Acc (%)
Teacher (Unpruned)	0	200	60.22
Standard FT [12]	4	200	61.58
Static KD	4	60	61.22
Dynamic CKA-KD (Ours)	**4**	**60**	**62.06**
Standard FT [12]	5	200	61.47
Static KD	5	60	61.81
Dynamic CKA-KD (Ours)	**5**	**60**	**61.81**
Standard FT [12]	9	200	60.92
Static KD	**9**	**60**	**61.70**
Dynamic CKA-KD (Ours)	9	60	61.59

Open-Set Face Recognition. The hyperspherical embedding space is exceptionally sensitive to structural disruptions. As shown in Table III, we evaluate the pruned iResNet-50 under a one-block pruning setting across LFW, CFP-FP, and AgeDB-30 using diverse recovery objectives and training subsets.

Experiments demonstrate that face recognition models exhibit severe vulnerability compared to closed-set classification. The ArcFace-only recovery collapses on cross-pose and age-variation benchmarks (e.g., reaching only 7.97% TAR on CFP-FP), as simultaneously optimizing the classification margin and alignment constraints post-pruning can create conflicting gradients.

TABLE III
TAR@FAR=1E-4 FOR 1-BLOCK-PRUNED FACE RECOGNITION MODELS (IRESNET-50).

Method / Training Setting	LFW	CFP-FP	AgeDB
Teacher (Unpruned)	**99.60**	**88.91**	**92.87**
ArcFace + Dynamic Align. (CASIA-1k)	87.73	31.74	7.80
ArcFace only (CASIA-1k)	83.77	7.97	11.43
Dynamic Align. only (CASIA-1k)	**99.37**	**80.17**	69.23
Dynamic Align. only (CASIA-3k)	**99.37**	78.00	70.93
Dynamic Align. only (MS1MV2-1k)	99.27	79.86	**72.07**

In contrast, utilizing Dynamic CKA-based alignment as the dominant recovery signal preserves the teacher's verification geometry far more effectively. With CASIA-1k, the student recovers 99.37% TAR on LFW and 80.17% on CFP-FP. Increasing identity diversity via the MS1MV2-1k subset yields the best AgeDB-30 result (72.07%), confirming that our dynamic weighting mechanism is essential for restoring angular structure and robustness.

While LFW recovery is near-perfect (99.37% vs. 99.60%), a noticeable gap persists on complex benchmarks like AgeDB-30 (72.07% vs. 92.87%). We attribute this to three factors. *First*, the ArcFace header must be randomly re-initialized since only backbone weights are inherited. Including this untrained header on a small subset (1k–3k IDs) injects high-variance gradients that distort the converged geometry; hence, pure feature distillation works best. *Second*, extreme data scarcity ($\sim$1% of MS1MV2) limits exposure to pose/age variations crucial for CFP-FP and AgeDB-30. *Third*, angular-margin embeddings are highly sensitive to structural perturbations: removing even one block shifts the hyperspherical manifold enough to degrade

strict FAR=10^{-4} thresholds. Together, these explain why frontal matching (LFW) recovers fully, while cross-pose/age tasks retain a gap.

To evaluate scalability under extreme compression, we further prune 20 residual blocks, compressing the ResNet-56 architecture by approximately 74%. Tables IV and V summarize the results on CIFAR-10 and CIFAR-100, respectively. On CIFAR-10, both the 200-epoch Standard FT (87.67%) and 60-epoch Static KD (87.64%) hit a hard ceiling, whereas our Dynamic CKA-KD reaches parity with the unpruned Teacher (trained via a standard protocol to purely isolate KD efficacy) at 87.92% in only 60 epochs. On CIFAR-100, Dynamic CKA-KD recovers 60.00% (vs. Teacher 60.22%), matching Static KD (60.01%) to within 0.01 pp while requiring only 30% of the baseline training budget. These results confirm that the CKA-guided recipe remains stable even when 74% of the network depth is removed. By recovering exact Teacher-level accuracy on CIFAR-10 and near-Teacher accuracy on CIFAR-100 under such severe compression, the framework provides a practical, ready-to-deploy solution for resource-constrained Edge AI environments.

TABLE IV
TOP-1 ACCURACY ON CIFAR-10 UNDER EXTREME PRUNING (20 BLOCKS PRUNED).

Method	Pruned Blks	Epochs	Acc (%)
Teacher (Unpruned)	0	200	87.92
Standard FT [12]	20	200	87.67
Static KD	20	60	87.64
Dynamic CKA-KD (Ours)	**20**	**60**	**87.92**

TABLE V
TOP-1 ACCURACY ON CIFAR-100 UNDER EXTREME PRUNING (20 BLOCKS PRUNED).

Method	Pruned Blks	Epochs	Acc (%)
Teacher (Unpruned)	0	200	60.22
Standard FT [12]	20	200	59.91
Static KD	**20**	**60**	**60.01**
Dynamic CKA-KD (Ours)	20	60	60.00

C. CKA Heatmap Interpretation

To empirically analyze the alignment behavior of our method, we visualize cross-network CKA heatmaps between the unpruned Teacher and the pruned Student on CIFAR-10 (Figure 2). Under Static KD, the heatmap is overwhelmingly diagonal and constrained, indicating blind layer-by-layer memorization that correlates with the representational collapse observed in Table I. Conversely, our Dynamic CKA-KD produces a broader, more distributed correlation signature, strongly suggesting that the Student learns a more independent, structurally optimized representation space governed by the "corridor of freedom," rather than merely copying the Teacher.

D. Training Efficiency and Convergence

Figure 3(a) illustrates the convergence advantage: under aggressive pruning, our Dynamic CKA-KD follows a steeper trajectory than the 200-epoch Standard FT baseline [12], reaching terminal accuracy in less than one-third of the training budget.

To quantify this resource advantage, we define the *Epoch Efficiency Score* ($\mathscr{S}_{\text{eff}}$) as the ratio between final accuracy and the number of epochs required for recovery ($\mathscr{S}_{\text{eff}} = \text{Accuracy}/\text{Total Epochs}$, which directly correlates with total wall-clock time, as detailed in our GitHub repository).

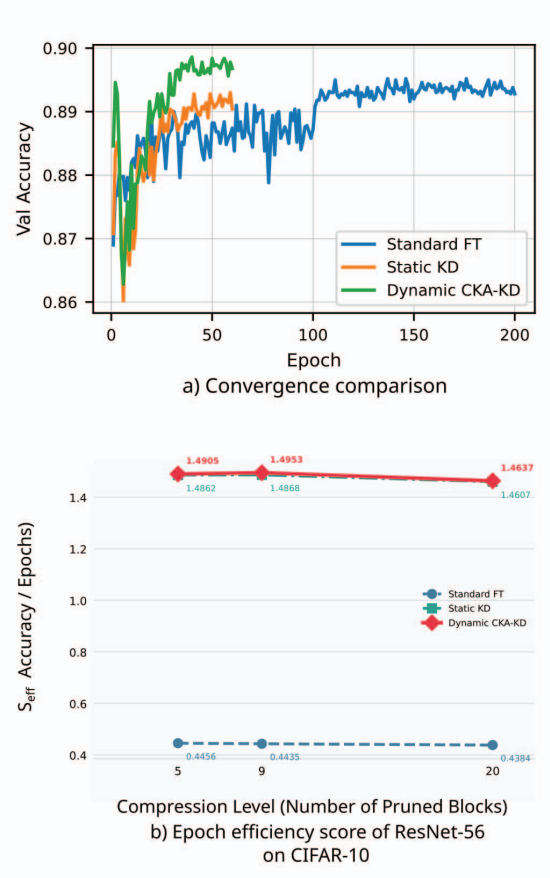

Fig. 3. Training efficiency and convergence behavior. (a) Validation accuracy convergence comparison under aggressive pruning, showing that Dynamic CKA-KD reaches stable recovery within 60 epochs compared with the 200-epoch Standard FT baseline. (b) Epoch Efficiency Score ($\mathscr{S}_{\text{eff}}$) across pruning levels, where the proposed method maintains a consistently higher accuracy-per-epoch trajectory.

As shown in Figure 3(b), the accelerated 60-epoch schedule yields an approximately 3.3× higher efficiency score than the 200-epoch baseline at the 9-block pruning milestone, while still producing superior accuracy. This pattern is preserved on CIFAR-100 and at the extreme 20-block setting (where the student recovers the Teacher accuracy of 87.92% despite a 74% depth reduction), confirming that Dynamic CKA-KD extracts substantially more accuracy per optimization epoch across datasets and compression levels.

E. Computational Cost Analysis

Table VI quantifies the computational savings achieved by block-level pruning on ResNet-56. Removing 20 blocks reduces the parameter count by 70.9% (from 861K to 251K) and FLOPs by 75.1% (from 125.7M to 31.4M), while CPU inference latency drops by 73.9% (from 1221.6 ms to 318.7 ms at batch size 128). These results demonstrate that the accuracy recovery reported in the preceding sections translates directly into substantial deployment-level efficiency gains.

TABLE VI
COMPUTATIONAL COST OF RESNET-56 UNDER INCREASING PRUNING DEPTH (BATCH SIZE 128, INTEL XEON @ 2.00 GHZ).

Pruned Blks	Params	FLOPs	CPU Latency (ms)
0 (Teacher)	861,770	125.75M	1221.56
5	795,978	102.15M	888.59
9	734,954	83.28M	682.59
20	250,922	31.38M	318.71

V. CONCLUSION AND FUTURE WORK

This paper addressed representational collapse in depth-pruned CNNs by introducing a distillation framework governed by Dynamic CKA-based Weighting. By establishing a representational "corridor of freedom," our mechanism empowers student models to optimize limited parameters with greater flexibility, bypassing counterproductive teacher mimicry. Coupled with an accelerated 60-epoch training recipe, this approach slashes conventional fine-tuning time by 70% while maintaining highly competitive baseline performance across rigorous classification and Face Recognition benchmarks. To ensure reproducibility, training logs and codes are available at https://github.com/DevinU-blue/Thesis-MultiStageKD-Dynamic-CKA. Building on this success, we plan to address the remaining gaps in complex face recognition tasks by integrating real-time CKA feedback into an autonomous auto-pruning framework. Furthermore, future extensions will investigate exhaustive ablation studies and multi-seed stability evaluations.

ACKNOWLEDGMENT

This work was supported by the University Research Grant from the University of Sciences, Hue University, under Project No. DHKH2026B-001.

REFERENCES

[1] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 770-778.

[2] I. C. Duta, L. Liu, F. Zhu and L. Shao, "Improved Residual Networks for Image and Video Recognition," 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 2021, pp. 9415-9422.

[3] M. Günther, S. Cruz, E. M. Rudd and T. E. Boult, "Toward Open-Set Face Recognition," 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, 2017, pp. 573-582.

[4] E. Caldeira, P. C. Neto, M. Huber, N. Damer, and A. F. Sequeira, "Model compression techniques in biometrics applications: A survey," Information Fusion, vol. 114, p. 102657, 2025, doi: 10.1016/j.inffus.2024.102657.

[5] C.-E. Wu, A. Davoodi, and Y. H. Hu, "Block pruning for enhanced efficiency in convolutional neural networks," arXiv (Cornell University), Dec. 2023, doi: 10.48550/arxiv.2312.16904.

[6] G. Li, K. Wang, P. Lv, P. He, Z. Zhou, and C. Xu, "Multistage feature fusion knowledge distillation," Scientific Reports, vol. 14, no. 1, p. 13373, Jun. 2024, doi: 10.1038/s41598-024-64041-4.

[7] J. Li, Z. Guo, H. Li, S. Han, J.-W. Baek, M. Yang, R. Yang, and S. Suh, "Rethinking feature-based knowledge distillation for face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 20156–20165..

[8] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj and L. Song, "SphereFace: Deep Hypersphere Embedding for Face Recognition," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 6738-6746, doi: 10.1109/CVPR.2017.713.

[9] S. Kornblith, M. Norouzi, H. Lee, and G. Hinton, "Similarity of neural network representations revisited," arXiv (Cornell University), May 2019, doi: 10.48550/arxiv.1905.00414.

[10] J. Liu et al., "UPDP: a unified progressive depth pruner for CNN and Vision Transformer," arXiv (Cornell University), Jan. 2024, doi: 10.48550/arxiv.2401.06426.

[11] F. Alonso-Fernandez, K. Hernandez-Diaz, J. M. B. Rubio, P. Tiwari, and J. Bigun, "Deep network pruning: A comparative study on CNNs in face recognition," Pattern Recognition Letters, vol. 189, pp. 221–228, Jan. 2025, doi: 10.1016/j.patrec.2025.01.023.

[12] I. Pons, B. Yamamoto, A. H. R. Costa, and A. Jordao, "Effective Layer Pruning Through Similarity Metric Perspective," arXiv (Cornell University), May 2024, doi: 10.48550/arxiv.2405.17081.

[13] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," arXiv (Cornell University), Mar. 2015, doi: 10.48550/arxiv.1503.02531.

[14] J. Tang et al., "Understanding and improving knowledge distillation," arXiv (Cornell University), Feb. 2020, doi: 10.48550/arxiv.2002.03532.

[15] L. Jing, P. Vincent, Y. LeCun, and Y. Tian, "Understanding dimensional collapse in contrastive self-supervised learning," arXiv (Cornell University), Oct. 2021, doi: 10.48550/arxiv.2110.09348.

[16] M. Davari, S. Horoi, A. Natik, G. Lajoie, G. Wolf, and E. Belilovsky, "Reliability of CKA as a similarity measure in deep learning," arXiv (Cornell University), Oct. 2022, doi: 10.48550/arxiv.2210.16156.

[17] J. Deng, J. Guo, N. Xue and S. Zafeiriou, "ArcFace: Additive Angular Margin Loss for Deep Face Recognition," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019, pp. 4685-4694, doi: 10.1109/CVPR.2019.00482.

[18] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "FitNets: Hints for Thin deep nets," arXiv (Cornell University), Dec. 2014, doi: 10.48550/arxiv.1412.6550.

[19] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," *arXiv preprint arXiv:1608.03983*, 2016.

[20] G. Fang *et al.*, "Depgraph: Towards any structural pruning," in *CVPR*, 2023, pp. 16091–16101.

[21] Y. Yang *et al.*, "Streamlined knowledge distillation," in *CVPR*, 2023, pp. 16120–16129.

[22] A. Krizhevsky, "Learning Multiple Layers of Features from Tiny Images," University of Toronto, Toronto, ON, Canada, Tech. Rep., 2012.

[23] G. B. Huang, M. Ramesh, T. Berg, and E. L. Miller, "Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments," University of Massachusetts Amherst, Amherst, MA, USA, Tech. Rep., Oct. 2007.

[24] S. Sengupta, J. -C. Chen, C. Castillo, V. M. Patel, R. Chellappa and D. W. Jacobs, "Frontal to profile face verification in the wild," 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Placid, NY, USA, 2016, pp. 1-9, doi: 10.1109/WACV.2016.7477558.

[25] S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia and S. Zafeiriou, "AgeDB: The First Manually Collected, In-the-Wild Age Database," 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, 2017, pp. 1997-2005, doi: 10.1109/CVPRW.2017.250.