Journal of Computer Networks, Architecture and High Performance Computing
Volume 08, Number 2, April 2026
https://doi.org/ 10.47709/cnahpc.v8i2.8239

Submitted : 2026-03-31
Accepted : 2026-04-21
Published : 2026-04-27

Hyperparameter Sensitivity of Vanilla Knowledge Distillation for Compact CNNs on CIFAR-100

Mochamad Rizal Fauzan[1)], Raden Muhammad Rafi Rachman[2)], Shifa Rangga Saputra[3)], Daffa Irsyad Nugraha[4)]
[1)]National Taipei University of Technology, Taipei 10608, Taiwan
[2,3,4)]Universitas Pendidikan Indonesia, Bandung 40154, Indonesia
[1)]rizalfauzan2002@email.com,[2)]rafirachmann@upi.edu, [3)]ranggaagasi@upi.edu,
[4)]dairnu22@upi.edu

ABSTRACT

Knowledge distillation has become an effective strategy for improving compact convolutional neural networks, yet the performance of vanilla knowledge distillation in lightweight image classification is still often reported using default hyperparameter settings without systematic justification. This study addresses the limited empirical understanding of how two core vanilla knowledge distillation hyperparameters, temperature scaling (**T**) and loss balancing (**α**), affect compact convolutional neural networks under a unified experimental setting. Using CIFAR-100 as the benchmark dataset, a ResNet-50 teacher was employed to distill knowledge into two lightweight student models, MobileNetV2 and ShuffleNetV2 ×1.0. Performance was evaluated using top-1 accuracy, top-5 accuracy, parameter count, and inference latency. The teacher achieved 81.24% top-1 accuracy and 96.05% top-5 accuracy. Under the default distillation setting, MobileNetV2 improved from 79.18% to 80.83% top-1 accuracy and from 95.77% to 96.40% top-5 accuracy, while reducing latency from 3.98 ms to 3.44 ms. ShuffleNetV2 ×1.0 improved from 77.00% to 78.36% top-1 accuracy and from 94.81% to 95.45% top-5 accuracy, with only a marginal latency increase from 4.23 ms to 4.29 ms. To examine hyperparameter sensitivity, an ablation study was conducted on MobileNetV2 with **T** = 2, 4, and 6, and **α** = 0.3, 0.5, and 0.7. The best configuration was obtained at **T** = 4 and **α** = 0.3, yielding 80.88% top-1 accuracy and 96.51% top-5 accuracy. These results show that vanilla knowledge distillation consistently improves compact convolutional neural networks, but its effectiveness depends strongly on careful hyperparameter selection rather than inherited default settings.

Keywords: CIFAR-100; compact neural networks; knowledge distillation; loss balancing; temperature scaling

INTRODUCTION

Deep convolutional neural networks have achieved remarkable performance in image classification, yet their deployment in real-world environments remains constrained by model size, memory consumption, and inference latency. These limitations are especially critical in resource-constrained scenarios, where a model is expected to preserve competitive predictive performance while maintaining efficient computation. To address this challenge, compact convolutional neural networks (CNNs) have been developed to reduce computational complexity without severely sacrificing accuracy. Among them, MobileNetV2 and ShuffleNetV2 are two representative lightweight architectures that are widely recognized for their efficiency-oriented design and practical suitability for edge and mobile deployment. MobileNetV2 emphasizes inverted residuals and linear bottlenecks, whereas ShuffleNetV2 is explicitly designed based on hardware-efficient operations and execution speed (Sandler et al., 2018; Ma et al., 2018). Although both

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 08, Number 2, April 2026
https://doi.org/ 10.47709/cnahpc.v8i2.8239

Submitted : 2026-03-31
Accepted : 2026-04-21
Published : 2026-04-27

architectures are intended for efficient inference, their distinct design principles suggest that they may respond differently to training strategies aimed at improving compact-model performance (C. Chen et al., 2025; Fauzan et al., 2025).

Knowledge distillation has been widely adopted as an effective compression-oriented training strategy because it allows a compact student model to learn not only from hard labels but also from the softened predictive distribution of a stronger teacher model. Through this mechanism, the student can exploit richer inter-class similarity information that is not directly available from one-hot supervision. As a result, knowledge distillation has become an attractive solution for improving lightweight models without changing their inference architecture. Compared with many accuracy-improvement methods that introduce additional computational modules or deployment overhead, distillation is particularly appealing because the additional complexity is largely confined to the training stage. This characteristic makes it well suited for compact CNNs that must operate under strict efficiency constraints (Y. Liu et al., 2024; Somantri et al., 2025).

Despite its popularity, the effectiveness of knowledge distillation is highly dependent on hyperparameter selection, particularly the temperature used to soften the logits and the coefficient that balances hard-target and soft-target supervision. In practice, however, these hyperparameters are often adopted from conventional defaults rather than justified through systematic evaluation. At the same time, a large portion of the recent literature has focused on proposing novel KD variants, modified transfer objectives, or increasingly sophisticated training schemes, while comparatively less attention has been paid to whether the standard vanilla KD formulation itself has already been properly configured for compact CNNs (Rybczak & Kozakiewicz, 2024). This creates a clear practical and methodological gap. If commonly reported KD gains are obtained under inherited hyperparameter choices, then part of the observed improvement may reflect arbitrary settings rather than well-understood knowledge transfer behavior. The issue becomes even more important for compact architectures, because different student designs may not benefit equally from the same distillation configuration (L. Liu et al., 2024). Unlike prior studies that focus on novel KD variants, this study systematically re-examines vanilla KD hyperparameters in compact CNN settings.

Based on this gap, this study investigates the hyperparameter sensitivity of vanilla knowledge distillation for compact CNNs in CIFAR-100 image classification. CIFAR-100 was selected because its multi-class structure provides a suitable benchmark for evaluating softened class-probability transfer in fine-grained recognition settings. A ResNet-50 model is used as the teacher, while MobileNetV2 and ShuffleNetV2 are used as student networks under a unified experimental protocol. The objective is to determine whether commonly used distillation settings remain appropriate for compact CNNs and to identify more suitable configurations when necessary. Specifically, this study makes three contributions. First, it provides a controlled comparison between standard supervised training and vanilla knowledge distillation on two representative lightweight architectures. Second, it conducts an ablation analysis of temperature scaling and loss balancing to examine how these two core hyperparameters influence compact student learning. Third, it evaluates the effect of distillation not only in terms of classification accuracy but also from an efficiency-oriented perspective that considers practical compact-model deployment.

LITERATURE REVIEW

Knowledge distillation was popularized by Hinton et al. as a teacher–student framework in which a compact model learns from softened class probabilities produced by a stronger teacher, allowing knowledge transfer beyond one-hot supervision. Since then, the literature has expanded from classical logit-based distillation to richer representational objectives. For example, Contrastive Representation Distillation reformulated transfer at the representation level and showed that standard KL-based distillation may overlook structural information embedded in teacher features, while Decoupled Knowledge Distillation

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 08, Number 2, April 2026
https://doi.org/ 10.47709/cnahpc.v8i2.8239

Submitted : 2026-03-31
Accepted : 2026-04-21
Published : 2026-04-27

revisited the original logit loss and separated target-class and non-target-class knowledge to improve the effectiveness of logit transfer. These developments confirm that the distillation literature has moved far beyond its initial formulation, with major attention given to increasingly sophisticated mechanisms for improving student learning (S. L. Chen et al., 2023; Zheng et al., 2023).

A second stream of research has emphasized that distillation performance depends not only on the loss formulation itself, but also on training dynamics and teacher–student compatibility. Cho and Hariharan showed that more accurate teachers do not always produce better students, highlighting the importance of architectural and capacity alignment in distillation. In a similar empirical spirit, Beyer et al. demonstrated that distillation can be highly effective when the training recipe is sufficiently patient and consistent, suggesting that implicit design choices may strongly influence the reported benefit of KD. More recently, DOT analyzed distillation from an optimization perspective and argued that conventional KD introduces a trade-off between task loss and distillation loss (Mao et al., 2024), which can limit convergence if the optimization process is not properly handled. Collectively, these studies indicate that KD effectiveness is not determined solely by whether distillation is applied, but also by how the training setup shapes the transfer process (Ma et al., 2023; Si et al., 2023; Zamanidoost et al., 2025).

Temperature scaling remains one of the most central yet least settled components of classification-based distillation. In the standard KD formulation, temperature controls the smoothness of teacher and student probability distributions and therefore directly affects how much relational information is exposed during training. However, recent works suggest that this design choice is still not fully understood. Asymmetric Temperature Scaling showed that conventional symmetric temperature scaling can reduce useful class discriminability when the teacher is much stronger than the student, motivating class-dependent temperature treatment. Curriculum Temperature for Knowledge Distillation further argued that a fixed temperature is often suboptimal and proposed a dynamic curriculum that adapts the distillation difficulty during training (Begum et al., 2024). Likewise, Transformed Teacher Matching re-examined temperature from a probabilistic perspective and reported that dropping temperature scaling on the student side can improve generalization. A very recent unified revisit also reported that the preferred temperature may depend on training configuration, teacher origin, student initialization, and dataset granularity. These findings strongly suggest that temperature should not be treated as a trivial default hyperparameter.

The relevance of these issues becomes even greater when the student network is a compact CNN intended for efficient deployment. MobileNetV2 and ShuffleNetV2 are both representative lightweight architectures, but they are built on different efficiency principles: MobileNetV2 relies on inverted residuals and linear bottlenecks, whereas ShuffleNetV2 is derived from practical guidelines that prioritize direct speed and memory-access efficiency rather than FLOPs alone. For such models, even modest changes in distillation behavior can meaningfully affect the final accuracy–efficiency trade-off. Recent empirical work has also shown that dataset characteristics can influence the benefit obtained from knowledge distillation in image classification, reinforcing the need for controlled evaluation rather than inherited defaults. Nevertheless, much of the recent literature continues to emphasize new KD variants, whereas fewer studies isolate the effect of vanilla KD hyperparameters such as temperature and loss balancing on compact convolutional students under a unified experimental protocol (J. Wang et al., 2023). Despite these advances, the effect of standard KD hyperparameters remains underexplored, especially in compact CNN settings where accuracy gains must be interpreted together with efficiency constraints. This is the specific gap addressed in this study.

METHOD

This study employed a quantitative experimental design to evaluate the effectiveness of knowledge distillation for compact convolutional neural networks in image classification. The experiment was designed to compare standard supervised training and teacher-guided distillation under a unified protocol, while also examining the effects of temperature scaling and loss balancing within the vanilla knowledge

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 08, Number 2, April 2026
https://doi.org/ 10.47709/cnahpc.v8i2.8239

Submitted : 2026-03-31
Accepted : 2026-04-21
Published : 2026-04-27

distillation framework. The overall workflow consisted of dataset preparation, teacher training, student baseline training, student distillation training, hyperparameter ablation, and final evaluation. The knowledge distillation setting followed the standard teacher student formulation, in which the student learns from both hard labels and softened probability distributions produced by the teacher. Figure 1 illustrates

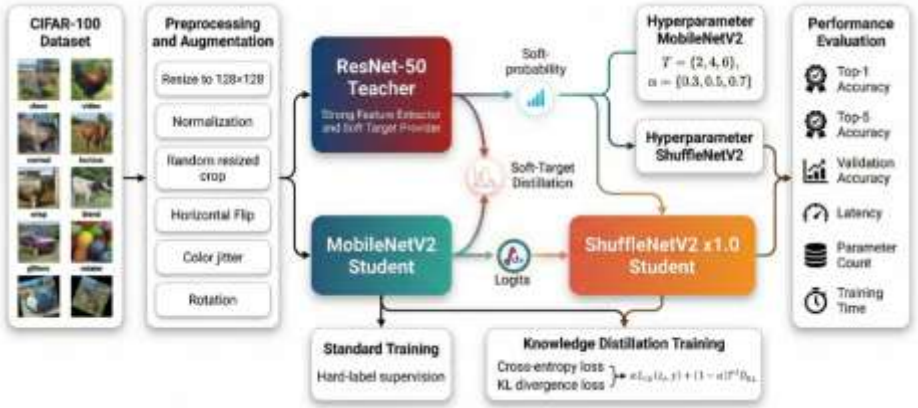


Figure 1. Experimental framework of the proposed knowledge distillation study.

the overall experimental framework of this study, including dataset preparation, teacher–student training, hyperparameter ablation, and final performance evaluation.

The experiments used the CIFAR-100 dataset (Krizhevsky & Hinton, 2009), which contains 100 image classes. The original training set contains 50,000 images and the test set contains 10,000 images. In this study, the original training set was further divided into 45,000 images for training and 5,000 images for validation using a fixed random split with seed 42. Because the selected backbones were initialized from ImageNet-pretrained weights, all images were resized from 32 × 32 to 128 × 128 pixels and normalized using ImageNet statistics. The training split used augmentation to improve generalization, while the evaluation split used deterministic preprocessing only. The full dataset and preprocessing configuration is summarized in Table 1.

Table 1. Dataset and preprocessing configuration

Component	Configuration
Dataset	CIFAR-100
Number of classes	100
Original training images	50,000
Training split	45,000
Validation split	5,000
Test images	10,000
Split strategy	Fixed random split with seed = 42
Original input size	32 × 32
Resized input size	128 × 128
Normalization mean	(0.485, 0.456, 0.406)
Normalization std	(0.229, 0.224, 0.225)
Training augmentation	Resize, RandomResizedCrop, RandomHorizontalFlip, ColorJitter, RandomRotation, Normalize
Evaluation preprocessing	Resize, CenterCrop, Normalize

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 08, Number 2, April 2026
https://doi.org/ 10.47709/cnahpc.v8i2.8239

Submitted : 2026-03-31
Accepted : 2026-04-21
Published : 2026-04-27

Three convolutional neural networks were used in the experiment. ResNet-50 served as the teacher model, while MobileNetV2 and ShuffleNetV2 ×1.0 were used as compact student models. ResNet-50 was selected to provide a stronger reference model, whereas MobileNetV2 and ShuffleNetV2 were selected as representative lightweight architectures with different efficiency-oriented design principles. All models were initialized using ImageNet-pretrained weights, and the final classifier layer of each network was modified to produce 100 output classes. The model configuration and the main optimization settings are presented in Table 2.

Table 2. Model and training configuration

Component	Configuration
Teacher model	ResNet-50
Student models	MobileNetV2; ShuffleNetV2 ×1.0
Teacher parameters	23.71 M
MobileNetV2 parameters	2.35 M
ShuffleNetV2 ×1.0 parameters	1.36 M
Teacher pretrained weights	ImageNet-1K V2
MobileNetV2 pretrained weights	ImageNet-1K V2
ShuffleNetV2 pretrained weights	ImageNet-1K V1
Teacher classifier modification	Linear(2048, 100)
MobileNetV2 classifier modification	Linear(1280, 100)
ShuffleNetV2 classifier modification	Linear(1024, 100)
Optimizer	AdamW
Learning rate	1 × 10-3
Weight decay	1 × 10-4
Scheduler	CosineAnnealingLR
Batch size	128
Epochs for main experiments	40
Epochs for ablation experiments	30
Gradient clipping	Max norm = 1.0
Precision setting	AMP with GradScaler
Random seed	42
DataLoader workers	0
cuDNN benchmark	Enabled

The training procedure was conducted in three stages. First, the teacher model was trained on the training split to provide the reference logits for the distillation process. Second, each student model was trained using standard supervised learning with cross-entropy loss to establish the non-distilled baseline. Third, each student model was retrained using knowledge distillation, in which the student simultaneously learned from ground-truth labels and softened teacher predictions. The distillation objective combined hard-target supervision and soft-target supervision. Let z_s and z_t denote the student and teacher logits, respectively, y denote the ground-truth label, T denote the temperature parameter, and α denote the loss-balancing coefficient. The total loss is defined as

$$\mathcal{L}_{KD} = \alpha \mathcal{L}_{CE}(z_s, y) + (1-\alpha) T^2 D_{KL}\left(\sigma\left(\frac{z_s}{T}\right) \parallel \sigma\left(\frac{z_t}{T}\right)\right) \quad (1)$$

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 08, Number 2, April 2026
https://doi.org/ 10.47709/cnahpc.v8i2.8239

Submitted : 2026-03-31
Accepted : 2026-04-21
Published : 2026-04-27

where $\mathcal{L}_{CE}$ is the cross-entropy loss, D_{KL} is the Kullback–Leibler divergence, and $\sigma(\cdot)$ is the softmax function. In the main distillation setting, the default configuration was $T = 4$ and $\alpha = 0.5$. To analyze hyperparameter sensitivity, an ablation study was performed using MobileNetV2 as the student and ResNet-50 as the teacher. Two controlled experiments were defined: temperature ablation with $T \in \{2,4,6\}$ while fixing $\alpha = 0.5$, and loss-balancing ablation with $\alpha \in \{0.3,0.5,0.7\}$ while fixing $T = 4$. All ablation runs used the same initialization strategy, dataset split, optimization setup, and training pipeline to preserve comparability. The distillation, ablation, and evaluation setup is summarized in Table 3.

Table 3. Distillation and evaluation setup

Component	Configuration
Distillation loss	$\alpha\mathcal{L}_{CE}(z_s, y) + (1-\alpha)T^2 D_{KL}$
Default temperature	$T = 4$
Default loss balance	$\alpha = 0.5$
Temperature ablation	$T = 2, 4, 6$ with $\alpha = 0.5$
Alpha ablation	$\alpha = 0.3, 0.5, 0.7$ with $T = 4$
Ablation student	MobileNetV2
Ablation teacher	ResNet-50
Checkpoint selection	Best validation top-1 accuracy
Evaluation metrics	Top-1 accuracy, Top-5 accuracy, best validation top-1, test loss, inference latency, training time
Latency protocol	Mean single-image inference time after warm-up and repeated measured iterations with GPU synchronization
Training-time protocol	Total wall-clock time including validation
Hardware	NVIDIA GeForce RTX 5060
Framework	PyTorch 2.11.0+cu130; torchvision 0.26.0+cu130
Software environment	Python 3.11.14 on Windows
Precision mode	Mixed precision (FP16/FP32)

The evaluation was designed to assess both predictive performance and deployment-oriented efficiency. Top-1 accuracy and top-5 accuracy were used to measure classification performance on the test set. Best validation top-1 accuracy was used to select the final checkpoint during training, while test loss was computed from the selected checkpoint on the held-out test set. Inference latency was measured as mean single-image execution time after a warm-up stage and repeated timed iterations with GPU synchronization, and training time was recorded as total wall-clock duration including validation. This evaluation design was intended to ensure that the compact student models were assessed not only by their predictive capability, but also by their practical suitability for efficient deployment.

RESULT

The results are presented in three parts, namely overall model comparison, knowledge distillation gain analysis, and hyperparameter ablation. The analysis emphasizes both predictive performance and deployment-oriented efficiency by considering top-1 accuracy, top-5 accuracy, validation performance, inference latency, and training time. Overall, knowledge distillation consistently improved both compact student architectures compared with standard supervised training. This pattern suggests that knowledge distillation improves compact CNNs consistently, but with diminishing returns as student architecture capacity becomes more constrained.

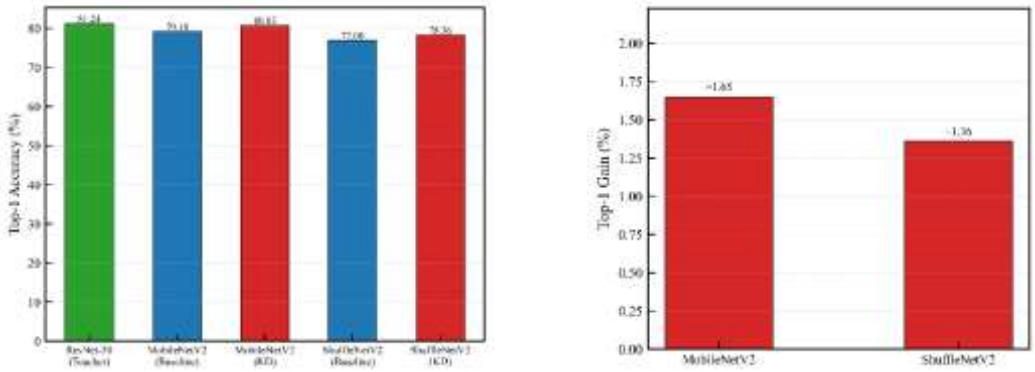
* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 08, Number 2, April 2026
https://doi.org/ 10.47709/cnahpc.v8i2.8239

Submitted : 2026-03-31
Accepted : 2026-04-21
Published : 2026-04-27

Overall Performance Comparison

Table 4 presents the overall comparison between the teacher model and the student models under standard and distillation-based training. ResNet-50 achieved the highest performance as the teacher, reaching 81.24% top-1 accuracy and 96.05% top-5 accuracy. Among the compact students, MobileNetV2 improved from 79.18% to 80.83% in top-1 accuracy and from 95.77% to 96.40% in top-5 accuracy after knowledge distillation. ShuffleNetV2 ×1.0 also improved from 77.00% to 78.36% in top-1 accuracy and from 94.81% to 95.45% in top-5 accuracy. In addition, both student models achieved higher best validation top-1 accuracy under knowledge distillation than under standard training, indicating that teacher-guided supervision improved the learning process in a consistent manner. Figure 2(a) visually confirms this pattern by showing that both distilled students moved closer to teacher-level performance. Figure 2(a) and Figure 2(b) further show that the distilled MobileNetV2 and ShuffleNetV2 maintained stronger validation behavior throughout training than their corresponding baselines.



(a) (b)

Figure 2. Overall accuracy improvement from knowledge distillation on CIFAR-100. (a) Top-1 accuracy comparison of teacher and student models. (b) Top-1 gain of compact student models after distillation.

Table 4. Overall performance comparison of teacher and student models.

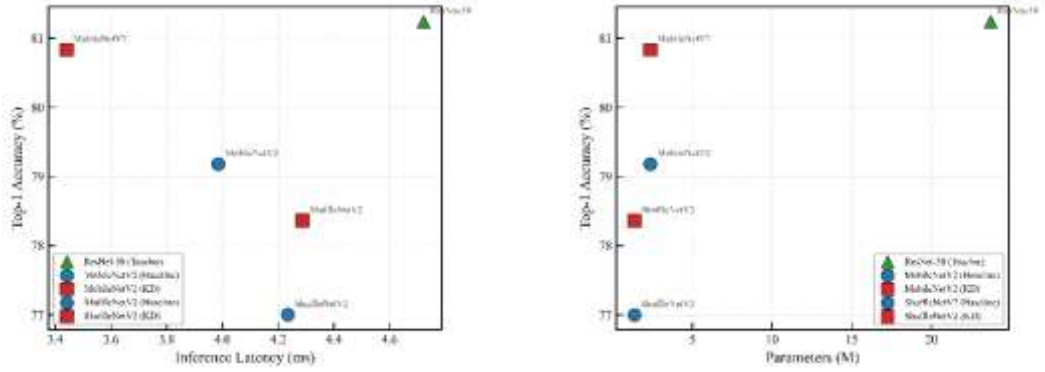
Model	Training Type	Params (M)	Best Val Top-1 (%)	Test Loss	Top-1 (%)	Top-5 (%)	Latency (ms)	Train Time (s)
ResNet-50	Teacher	23.71	82.04	0.9568	81.24	96.05	4.72	5677.43
MobileNetV2	Standard	2.35	80.32	1.0382	79.18	95.77	3.98	5304.86
MobileNetV2	KD	2.35	81.56	1.0814	80.83	96.40	3.44	5471.55
ShuffleNetV2 ×1.0	Standard	1.36	76.82	1.0456	77.00	94.81	4.23	5006.53
ShuffleNetV2 ×1.0	KD	1.36	79.04	1.2488	78.36	95.45	4.29	5308.46

From an efficiency perspective, MobileNetV2 under knowledge distillation provided the strongest compact-model result because it reached 80.83% top-1 accuracy while reducing latency from 3.98 ms to 3.44 ms. ShuffleNetV2 ×1.0 also benefited from distillation with a clear accuracy gain and only a marginal latency increase from 4.23 ms to 4.29 ms. Figure 4(a) and Figure 4(b) illustrate these efficiency-oriented relationships more clearly by positioning the teacher and student models in the accuracy–latency and accuracy–parameter spaces, respectively. These results indicate that knowledge distillation improved compact CNN performance without modifying the student architecture during inference.

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 08, Number 2, April 2026
https://doi.org/ 10.47709/cnahpc.v8i2.8239

Submitted : 2026-03-31
Accepted : 2026-04-21
Published : 2026-04-27

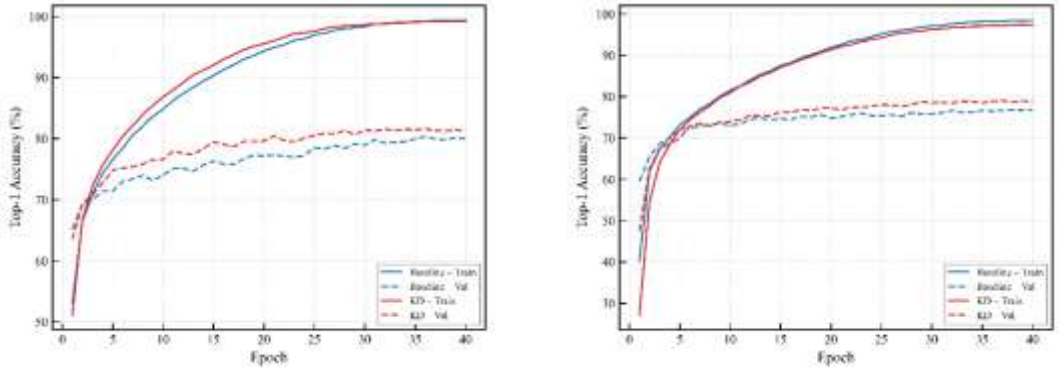


(a) (b)

Figure 3. Efficiency comparison of teacher and student models. (a) Top-1 accuracy versus inference latency. (b) Top-1 accuracy versus parameter count.

Knowledge Distillation Gain Analysis

To provide a clearer view of the effect of knowledge transfer, Table 5 summarizes the performance gains achieved by each student architecture. MobileNetV2 gained 1.65 percentage points in top-1 accuracy and 0.63 percentage points in top-5 accuracy. ShuffleNetV2 ×1.0 gained 1.36 percentage points in top-1 accuracy and 0.64 percentage points in top-5 accuracy. These gains demonstrate that the benefit of distillation was not restricted to a single lightweight design, but was observed across two compact architectures with different computational principles. Figure 2(b) complements this result by directly visualizing the top-1 improvement contributed by distillation for each student model.



(a) (b)

Figure 4. Training and validation convergence of compact student models. (a) MobileNetV2 under standard training and knowledge distillation. (b) ShuffleNetV2 ×1.0 under standard training and knowledge distillation.

Table 5. Knowledge distillation gain per student architecture

Student Model	Standard Top-1 (%)	KD Top-1 (%)	Top-1 Gain (%)	Standard Top-5 (%)	KD Top-5 (%)	Top-5 Gain (%)	Latency Difference (ms)
MobileNetV2	79.18	80.83	+1.65	95.77	96.40	+0.63	-0.54
ShuffleNetV2 ×1.0	77.00	78.36	+1.36	94.81	95.45	+0.64	+0.05

The gain pattern also suggests that MobileNetV2 benefited slightly more from teacher guidance than

* Corresponding author

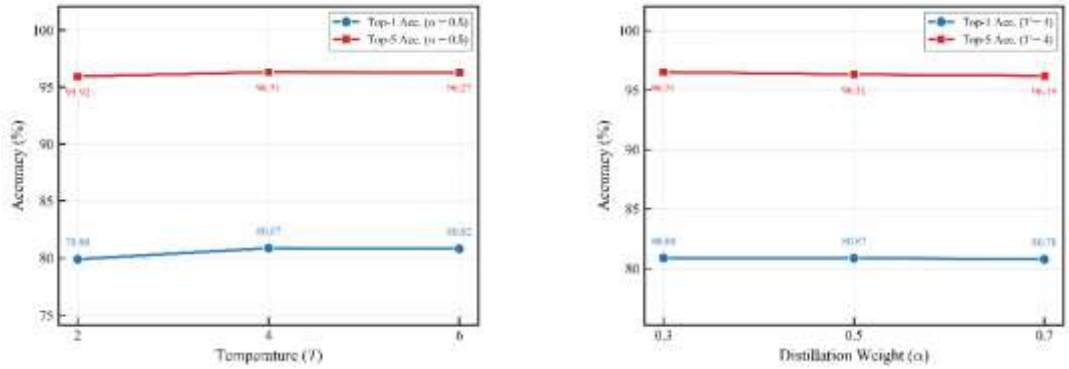
Journal of Computer Networks, Architecture and High Performance Computing
Volume 08, Number 2, April 2026
https://doi.org/ 10.47709/cnahpc.v8i2.8239

Submitted : 2026-03-31
Accepted : 2026-04-21
Published : 2026-04-27

ShuffleNetV2 ×1.0 in terms of top-1 improvement. At the same time, the top-5 gains of both models were highly similar, indicating that knowledge distillation consistently improved class ranking performance across compact students. The stronger gain observed on MobileNetV2 may indicate that its larger parameter budget enabled more effective absorption of the teacher's softened knowledge.

Hyperparameter Ablation Results

Table 6 summarizes the ablation study conducted on MobileNetV2 to analyze the sensitivity of vanilla knowledge distillation to temperature scaling and loss balancing. When the loss-balancing coefficient was fixed at $\alpha = 0.5$, the best test top-1 accuracy was obtained at $T = 4$, reaching 80.87%, whereas $T = 6$ produced the highest validation accuracy of 81.60% but slightly lower test accuracy. When the temperature was fixed at $T = 4$, the best overall result was achieved at $\alpha = 0.3$, which yielded 80.88% top-1 accuracy and 96.51% top-5 accuracy. As α increased from 0.3 to 0.7, both top-1 and top-5 accuracy decreased gradually. Figure 5(a) and Figure 5(b) visually reinforce these trends by showing that moderate temperature and lower hard-label weighting provided the most favorable distillation setting in this experiment.



(a) (b)

Figure 5. Hyperparameter sensitivity of vanilla knowledge distillation on MobileNetV2. (a) Effect of temperature scaling with fixed $\alpha = 0.5$. (b) Effect of loss balancing with fixed $T = 4$.

Table 6. Ablation results for temperature scaling and loss balancing on MobileNetV2

Ablation Type	Setting	Best Val Top-1 (%)	Top-1 (%)	Top-5 (%)
Temperature	$T = 2, \alpha = 0.5$	80.26	79.90	95.92
Temperature	$T = 4, \alpha = 0.5$	81.08	80.87	96.31
Temperature	$T = 6, \alpha = 0.5$	81.60	80.82	96.27
Loss balancing	$\alpha = 0.3, T = 4$	81.36	80.88	96.51
Loss balancing	$\alpha = 0.5, T = 4$	81.08	80.87	96.31
Loss balancing	$\alpha = 0.7, T = 4$	81.02	80.78	96.19

The ablation results show that the effectiveness of vanilla knowledge distillation was not determined solely by the teacher–student framework itself, but also by the choice of hyperparameters. A very low temperature reduced the benefit of soft-target transfer, whereas a higher temperature improved validation performance but did not yield the strongest final test accuracy. Likewise, a lower α value consistently produced better results than higher α settings, suggesting that stronger emphasis on the soft-target component was more beneficial for the evaluated compact student. Taken together, these findings confirm that careful hyperparameter selection is necessary to obtain the full benefit of knowledge distillation in compact convolutional neural networks.

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 08, Number 2, April 2026
https://doi.org/ 10.47709/cnahpc.v8i2.8239

Submitted : 2026-03-31
Accepted : 2026-04-21
Published : 2026-04-27

DISCUSSIONS

The results confirm that vanilla knowledge distillation remains effective for improving compact convolutional neural networks in image classification. Both student models benefited from teacher-guided learning, as MobileNetV2 and ShuffleNetV2 ×1.0 achieved higher top-1 accuracy after distillation than under standard supervised training. From a knowledge transfer perspective, this finding supports the central premise of distillation that softened teacher predictions convey informative inter-class relationships beyond hard labels alone (Y. Wang et al., 2024). In other words, the student is guided not only toward the correct class, but also toward the relative similarity structure learned by the teacher, which improves the quality of supervision under limited model capacity. This interpretation is consistent with the view of knowledge distillation as an information transfer mechanism, where the teacher provides a richer supervisory signal than one-hot targets. At the same time, the present results indicate that the effectiveness of distillation should not be treated merely as a binary question of whether KD works, but rather as a question of how effectively that transferred information is shaped by hyperparameter selection.

The ablation analysis strengthens this interpretation from both optimization and generalization perspectives. The temperature study showed that $T = 4$ produced the best test performance, whereas a higher temperature yielded slightly better validation accuracy but weaker final generalization. From an information-transfer standpoint, temperature determines how much relational structure is revealed in the teacher distribution: if T is too low, the output remains overly sharp and provides limited dark knowledge, whereas if T is too high, the distribution may become excessively smooth and lose discriminative value. This pattern can also be interpreted through the bias-variance tradeoff. A moderate temperature appears to provide sufficient regularization to reduce variance during student learning, while still preserving class discrimination and preventing excessive bias. A similar trend was observed in the loss-balancing analysis, where $\alpha = 0.3$ achieved the best result and larger α values gradually reduced performance. Since α controls the relative contribution of hard-label supervision, these results suggest that stronger emphasis on soft-target learning produced a better optimization balance for the evaluated compact student. In this sense, vanilla KD can be understood as a multi-objective optimization problem in which the student must simultaneously fit the ground-truth labels and align with the teacher's softened output distribution. The best-performing setting was therefore not the one that maximized either objective in isolation, but the one that produced the most favorable balance between task fidelity and transferable teacher information (Rafidison et al., 2023; Song et al., 2025).

The architecture-dependent gain pattern also provides an important theoretical implication. MobileNetV2 benefited slightly more from distillation than ShuffleNetV2 ×1.0, suggesting that compact models do not absorb transferred knowledge equally. This may reflect differences in representational capacity and optimization flexibility: a student with slightly greater effective capacity may be better able to exploit the additional structure contained in teacher outputs, whereas a more constrained architecture may experience diminishing returns even under the same teacher signal. From an efficiency perspective, however, both students improved accuracy without changing their inference architecture, which confirms the practical relevance of KD for deployment-oriented compact CNNs. Nevertheless, this study has several limitations. The experiments were conducted only on CIFAR-100, used a single teacher architecture, and performed hyperparameter ablation only on MobileNetV2. In addition, the study focused on vanilla knowledge distillation, used a single random seed, and resized CIFAR-100 inputs, which may limit generalizability and direct comparability with other settings. Therefore, future work should extend the evaluation to multiple datasets, multiple seeds, additional teacher models, and more advanced distillation methods to determine whether the observed optimization and information-transfer patterns remain consistent across broader compact-model settings.

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 08, Number 2, April 2026
https://doi.org/ 10.47709/cnahpc.v8i2.8239

Submitted : 2026-03-31
Accepted : 2026-04-21
Published : 2026-04-27

CONCLUSION

This study revisited vanilla knowledge distillation for compact convolutional neural networks by analyzing the roles of temperature scaling and loss balancing in CIFAR-100 image classification. The results showed that knowledge distillation consistently improved both evaluated student architectures under the same training protocol. MobileNetV2 improved from 79.18% to 80.83% top-1 accuracy and from 95.77% to 96.40% top-5 accuracy, while ShuffleNetV2 ×1.0 improved from 77.00% to 78.36% top-1 accuracy and from 94.81% to 95.45% top-5 accuracy. These findings confirm that teacher-guided soft-target supervision remains an effective strategy for strengthening compact convolutional models without modifying their inference architecture. In addition, the study showed that the practical value of knowledge distillation should be assessed not only from predictive accuracy but also from an efficiency perspective, since the distilled compact models preserved lightweight deployment characteristics while achieving measurable performance gains.

The ablation analysis further demonstrated that the effectiveness of vanilla knowledge distillation is strongly influenced by hyperparameter selection. For MobileNetV2, the best temperature setting in the evaluated configuration was $T = 4$, while the best loss-balancing configuration was $\alpha = 0.3$, which achieved 80.88% top-1 accuracy and 96.51% top-5 accuracy. These results indicate that moderate temperature scaling and stronger emphasis on soft-target supervision provide the most favorable setting for the evaluated compact student. More importantly, this study challenges the common practice of using default KD settings without systematic validation, and shows that careful hyperparameter selection is necessary to obtain more reliable and practically meaningful gains in compact CNNs. This is important because reported distillation improvements may otherwise reflect inherited experimental defaults rather than well-optimized knowledge transfer behavior. Therefore, the main contribution of this study lies not only in reporting accuracy gains, but also in demonstrating that re-examining vanilla KD hyperparameters is essential for fairer evaluation and better deployment-oriented model optimization. Future work should extend this analysis to additional datasets, multiple random seeds, different teacher architectures, and more advanced distillation variants to further validate the generality of the observed trends.

REFERENCES

Begum, M., Hasan Shuvo, M., Kamal Nasir, M., Hossain, A., Jakir Hossain, M., Ashraf, I., Uddin, J., & Samad, M. A. (2024). LCNN: Lightweight CNN Architecture for Software Defect Feature Identification Using Explainable AI. *IEEE Access*, *12*(April), 55744–55756. https://doi.org/10.1109/ACCESS.2024.3388489

Chen, C., Mat Isa, N. A., & Liu, X. (2025). A review of convolutional neural network based methods for medical image classification. *Computers in Biology and Medicine*, *185*, 109507. https://doi.org/10.1016/j.compbiomed.2024.109507

Chen, S. L., Chen, T. Y., Mao, Y. C., Lin, S. Y., Huang, Y. Y., Chen, C. A., Lin, Y. J., Chuang, M. H., & Abu, P. A. R. (2023). Detection of Various Dental Conditions on Dental Panoramic Radiography Using Faster R-CNN. *IEEE Access*, *11*(November), 127388–127401. https://doi.org/10.1109/ACCESS.2023.3332269

Fauzan, M. R., Pramudita, R., Rizqulloh, M. A., & Sartika, N. (2025). Integrated Energy Monitoring and Control System with Tri-Node ESP32 Architecture. *Proceedings of 2025 11th International Conference on Wireless and Telematics, ICWT 2025*, 1–6. https://doi.org/10.1109/ICWT66752.2025.11181758

Krizhevsky, A., & Hinton, G. (2009). *Learning Multiple Layers of Features from Tiny Images*. https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf

Liu, L., Wang, Y., Peng, J., & Zhang, L. (2024). GLR-CNN: CNN-Based Framework With Global Latent Relationship Embedding for High-Resolution Remote Sensing Image Scene Classification. *IEEE*

* Corresponding author

Journal of Computer Networks, Architecture and High Performance Computing
Volume 08, Number 2, April 2026
https://doi.org/ 10.47709/cnahpc.v8i2.8239

Submitted : 2026-03-31
Accepted : 2026-04-21
Published : 2026-04-27

Transactions on Geoscience and Remote Sensing, *62*, 1–13. https://doi.org/10.1109/TGRS.2024.3434452

Liu, Y., Xue, J., Li, D., Zhang, W., Chiew, T. K., & Xu, Z. (2024). Image recognition based on lightweight convolutional neural network: Recent advances. *Image and Vision Computing*, *146*, 105037. https://doi.org/10.1016/j.imavis.2024.105037

Ma, N., Sun, L., He, Y., Zhou, C., & Dong, C. (2023). CNN-TransNet: A Hybrid CNN-Transformer Network With Differential Feature Enhancement for Cloud Detection. *IEEE Geoscience and Remote Sensing Letters*, *20*, 1–5. https://doi.org/10.1109/LGRS.2023.3288742

Ma, N., Zhang, X., Zheng, H. T., & Sun, J. (2018). Shufflenet V2: Practical guidelines for efficient cnn architecture design. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11218 LNCS*, 122–138. https://doi.org/10.1007/978-3-030-01264-9_8

Mao, S., Li, H., Zhang, Y., & Shi, Y. (2024). Prediction of Ionospheric Electron Density Distribution Based on CNN-LSTM Model. *IEEE Geoscience and Remote Sensing Letters*, *21*, 1–5. https://doi.org/10.1109/LGRS.2024.3437650

Rafidison, M. A., Ramafiarisona, H. M., Randriamitantsoa, P. A., Rafanantenana, S. H. J., Toky, F. M. R., Rakotondrazaka, L. P., & Rakotomihamina, A. H. (2023). Image Classification Based on Light Convolutional Neural Network Using Pulse Couple Neural Network. *Computational Intelligence and Neuroscience*, *2023*(1), 7371907. https://doi.org/10.1155/2023/7371907

Rybczak, M., & Kozakiewicz, K. (2024). Deep Machine Learning of MobileNet, Efficient, and Inception Models. *Algorithms 2024, Vol. 17, Page 96*, *17*(3), 96. https://doi.org/10.3390/a17030096

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). MobileNetV2: Inverted Residuals and Linear Bottlenecks. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 4510–4520. https://doi.org/10.1109/CVPR.2018.00474

Si, M., Wang, Y., Siljak, H., Seow, C., & Yang, H. (2023). A Lightweight CIR-Based CNN With MLP for NLOS/LOS Identification in a UWB Positioning System. *IEEE Communications Letters*, *27*(5), 1332–1336. https://doi.org/10.1109/LCOMM.2023.3260953

Somantri, M., Fauzan, M. R., & Surya, I. (2025). Optimization of IoT-based monitoring system for automatic power factor correction using PZEM-004T sensor. *Indonesian Journal of Electrical Engineering and Computer Science*, *39*(2), 860. https://doi.org/10.11591/ijeecs.v39.i2.pp860-873

Song, J., Liang, R., Yuan, B., & Hu, J. (2025). DiMO-CNN: Deep Learning Toolkit-Accelerated Analytical Modeling and Optimization of CNN Hardware and Dataflow. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, *44*(1), 251–265. https://doi.org/10.1109/TCAD.2024.3429419

Wang, J., Zhang, X., Gao, G., Lv, Y., Li, Q., Li, Z., Wang, C., & Chen, G. (2023). Open Pose Mask R-CNN Network for Individual Cattle Recognition. *IEEE Access*, *11*(September), 113752–113768. https://doi.org/10.1109/ACCESS.2023.3321152

Wang, Y., Zhang, T., Zhao, L., Hu, L., Wang, Z., Niu, Z., Cheng, P., Chen, K., Zeng, X., Wang, Z., Wang, H., & Sun, X. (2024). RingMo-Lite: A Remote Sensing Lightweight Network With CNN-Transformer Hybrid Framework. *IEEE Transactions on Geoscience and Remote Sensing*, *62*, 1–20. https://doi.org/10.1109/TGRS.2024.3360447

Zamanidoost, Y., Ould-Bachir, T., & Martel, S. (2025). OMS-CNN: Optimized Multi-Scale CNN for Lung Nodule Detection Based on Faster R-CNN. *IEEE Journal of Biomedical and Health Informatics*, *29*(3), 2148–2160. https://doi.org/10.1109/JBHI.2024.3507360

Zheng, C., Hu, C., Chen, Y., & Li, J. (2023). A Self-Learning-Update CNN Model for Semantic Segmentation of Remote Sensing Images. *IEEE Geoscience and Remote Sensing Letters*, *20*, 1–5. https://doi.org/10.1109/LGRS.2023.3261402

* Corresponding author