Knowledge Distillation on YOLOv8

Arav Adikesh Ramakrishnan
aravadikeshr@umass.edu

Siddhartha Jaiswal
sjaiswal@umass.edu

Abstract

Knowledge distillation represents a critical approach for developing lightweight, efficient machine learning models that can perform comparably to larger, more complex architectures. This study introduces an advanced knowledge distillation methodology for YOLOv8-based object classification models, leveraging logit standardization and curriculum temperature scheduling techniques. By transferring knowledge from a large YOLOv8l teacher model to a compact YOLOv8n student model, we demonstrate significant performance improvements across challenging benchmark datasets. Our experimental results reveal a 5%*,* 35.36% *and* 37.61% *relative increase in classification accuracy on CIFAR-10, Tiny ImageNet and Oxford-IIIT-Pet respectively, without compromising inference time. Comprehensive evaluations across CIFAR-10, Tiny ImageNet, and the Oxford-IIIT-Pet datasets validate the proposed approach's robustness and generalizability. This research contributes to the growing field of model optimization, offering a promising strategy for developing efficient machine learning solutions for resource-constrained environments.*

1. Introduction

Knowledge distillation has emerged as a transformative methodology for transferring sophisticated representations from large, complex machine learning models to smaller, more computationally efficient architectures. This approach addresses a critical challenge in modern artificial intelligence: developing high-performance models that can operate effectively within stringent computational and resource constraints. By strategically mimicking the intricate output distributions and intermediate representations of a sophisticated "teacher" model, smaller "student" models can achieve performance levels approaching their larger counterparts while dramatically reducing computational overhead, memory requirements, and energy consumption. The technique is particularly compelling in computer vision classification tasks, where maintaining high accuracy is paramount, yet deployment frequently occurs in resource-limited environments such as edge devices, mobile platforms, autonomous systems, and specialized scientific instrumentation. Traditional deep learning approaches often require substantial computational resources, making them impractical for deployment in scenarios with limited processing power, constrained memory, or strict energy efficiency requirements. Knowledge distillation offers an elegant solution to this fundamental optimization challenge, enabling the development of compact, high-performance models that can operate effectively across diverse computational contexts.

In this study, we explore advanced knowledge distillation techniques specifically optimized for object classification models within the YOLOv8 [3] architectural framework. YOLO (You Only Look Once) models have garnered significant attention in computer vision research for their exceptional balance of speed and accuracy, with the YOLOv8 series offering specialized variants designed for diverse application domains. Our research strategically leverages a high-capacity YOLOv8-Large model (YOLOv8l) as the teacher, meticulously transferring its learned representational knowledge to a more compact YOLOv8-Nano variant (YOLOv8n) fine-tuned for classification tasks. To maximize the efficacy of knowledge transfer, we introduce two sophisticated distillation techniques: logit standardization and curriculum temperature scheduling. These innovative approaches are engineered to stabilize and refine the knowledge transfer process, ensuring precise and meaningful information transmission between model architectures while mitigating potential representational degradation.

Our experimental results demonstrate significant performance enhancements for the YOLOv8 nano model, including a statistically significant 5% increase in classification accuracy on the CIFAR-10 dataset, achieved without compromising inference time or computational overhead. By implementing these advanced distillation techniques, we enable the lightweight student model to effectively navigate fine-grained classification challenges and complex real-world scenarios without substantial accuracy trade-offs.

2. Related Work

Knowledge distillation (KD) has emerged as a critical paradigm for model compression and performance optimization, evolving from traditional logit-based approaches to sophisticated multi-modal knowledge transfer techniques. This research synthesizes advanced logit-based methodologies to enhance the performance of lightweight object classification models.

2.1. Logit-Based Distillation

Pioneered by Hinton et al. [2], logit-based distillation represents a fundamental approach to knowledge transfer between neural network architectures. The core mechanism involves transforming teacher model logits through temperature-based softmax scaling, which enables more nuanced probability distributions and facilitates the student model's learning of complex inter-class relationships. Recent advancements have significantly refined this foundational approach, introducing two critical techniques that address inherent limitations in traditional knowledge distillation:

2.2. Logit Standardization

Proposed by Sun et al. [8], logit standardization addresses the fundamental challenge of logit variability during knowledge transfer. By implementing a normalization process before softmax transformation, this technique ensures:

More uniform contribution of individual logits to the final probability distribution Reduced susceptibility to high-variance logits that might introduce computational noise Enhanced stability in knowledge transfer, particularly in complex, fine-grained classification scenarios

The standardization process effectively mitigates potential information distortion, creating a more robust knowledge transfer mechanism that preserves the intricate representational learning of the teacher model.

2.3. Curriculum Temperature Knowledge Distillation (CT-KD)

Li et al. [6] introduced a dynamic temperature scheduling approach that fundamentally reimagines the knowledge distillation learning process. Unlike static temperature-based methods, CT-KD implements a progressive learning strategy characterized by:

Initial high-temperature phases that facilitate smoother, more generalized learning Gradual temperature reduction to enable increasingly precise feature discrimination A curriculum-like learning mechanism that mimics natural cognitive skill acquisition

This approach recognizes that knowledge transfer is not a uniform process but a nuanced progression from broad conceptual understanding to refined, task-specific representations.

2.4. Synergistic Knowledge Transfer

The combined application of logit standardization and curriculum temperature scheduling represents a sophisticated approach to knowledge distillation. By addressing both the statistical properties of logit distributions and the dynamic learning trajectory of the student model, these techniques offer a comprehensive strategy for efficient model compression and performance optimization.

3. Methodology

3.1. YOLO Architecture for Classification

The YOLO framework, originally pioneered for object detection, has been strategically adapted for image classification tasks through architectural modifications and specialized training objectives. Given an input image I, the YOLO classification model generates a probabilistic mapping $p = f(I; \theta)$, where f represents the neural network transformation and θ denotes the model's learned parameters.

The classification pipeline comprises three critical stages:

- **Hierarchical Feature Extraction:** The input image traverses through a multi-scale convolutional backbone network, systematically extracting hierarchical spatial and semantic representations. This process captures increasingly abstract features, from low-level edge and texture information to high-level semantic concepts.
- **Feature Aggregation:** Extracted features undergo global average pooling, a dimensionality reduction technique that condenses spatial feature maps into a compact, semantically rich feature vector. This operation effectively captures the global context while maintaining computational efficiency.
- **Probabilistic Classification:** The aggregated feature vector is processed through fully connected layers with a softmax activation, generating a normalized probability distribution across predefined class categories.

The model's optimization leverages cross-entropy loss, formally defined as:

$$\mathcal{L}_{\text{CE}} = -\sum_{c=1}^{C} y_c \log(p_c), \tag{1}$$

where y_c represents the ground truth binary indicator for class c and p_c denotes the model's predicted probability. This loss function incentivizes the model to maximize the likelihood of correct class prediction.

Our research introduces a specialized modification to the standard YOLOv8 architecture to optimize performance for fine-grained classification tasks, particularly on datasets like CIFAR-10. This adaptation addresses critical chal-

lenges in transferring object detection architectures to classification domains.

3.1.1 Architectural Modifications

The modified YOLOv8 model undergoes a systematic restructuring to enhance its classification capabilities:

- **Backbone Preservation:** The original convolutional backbone is retained, preserving the hierarchical feature extraction capabilities of the YOLOv8 architecture. Copy
- **Feature Dimensionality Handling:** A dynamic feature extraction mechanism determines the optimal number of feature dimensions from the final convolutional layer, specifically the C2f layer's cv2 convolution.
- **Classification Head Redesign:** The original detection head is replaced with a streamlined classification pipeline comprising: - Adaptive average pooling to standardize spatial features - Flattening of feature representations - Dropout regularization - A final linear classification layer

3.2. Knowledge Distillation Formulation

The knowledge distillation loss integrates two complementary loss components:

1. **Soft Target Loss:** Kullback-Leibler (KL) divergence between temperature-scaled logits, capturing nuanced inter-class relationships.
2. **Hard Target Loss:** Standard cross-entropy loss with ground truth labels, ensuring direct supervised learning.

The comprehensive loss function is formally expressed as:

$$\mathcal{L} = (1-\alpha)\mathcal{L}_{\text{CE}}(y, \sigma(z^s)) + \alpha T^2 \mathcal{L}_{\text{KL}}(q^t \| q^s) \quad (2)$$

where α modulates the contribution of soft and hard targets, T represents the temperature scaling parameter, and σ denotes the softmax activation.

3.3. Logit Standardization

Following the methodology proposed by Sun et al. [8], logit standardization is implemented through:

$$\hat{z} = \frac{z - \mu_z}{\sigma_z + \epsilon} \quad (3)$$

This normalization mitigates logit variance, promoting more stable and informative knowledge transfer by ensuring uniform contribution across feature dimensions.

3.4. Curriculum Temperature Scheduling

The temperature parameter evolves dynamically throughout training:

$$T(e) = \max(1.0, T_0 \cdot \gamma^e) \quad (4)$$

This adaptive scheduling facilitates a progressive learning regime: initial high-temperature phases enable broad feature exploration, while gradually reduced temperatures encourage refined, discriminative feature learning.

4. Experimentation Details

4.1. Model Details and Pre-Training

Dataset The experiments conducted in this study leverage the CIFAR-10 [5] and Tiny-ImageNet [1] datasets to pre-train and evaluate the performance of YOLOv8-based teacher and student models. CIFAR-10 is a well-established dataset comprising 60,000 images across 10 classes, offering a relatively simpler classification task. In contrast, Tiny-ImageNet, which includes 100,000 images from 200 classes, provides a more challenging and diverse dataset. This combination of datasets allows for a comprehensive assessment of the models' abilities to handle standard classification tasks.

For the teacher model, YOLOv8Large (YOLOv8l) was selected due to its robust architecture, high capacity, and strong baseline performance. For the student model, YOLOv8Nano (YOLOv8n) was chosen, as it is a lightweight variant designed for resource-constrained environments. Table 1 summarize the baseline performance of both models on CIFAR-10 and Tiny-ImageNet, respectively.

As illustrated in the tables, the teacher model, YOLOv8Large, significantly outperforms the student model, YOLOv8Nano, across both datasets in terms of accuracy. However, these gains come at the cost of higher inference times and increased computational complexity, as indicated by the number of parameters and floating-point operations (FLOPs). The high accuracy of YOLOv8Large underscores its potential as a strong teacher model for knowledge distillation, while the efficiency of YOLOv8Nano highlights its suitability for deployment in real-time and resource-constrained scenarios.

Evaluation metrics. We employ top-1 accuracy, precision, recall@1 and the F1 score to comprehensively evaluate model performance.

Training details. The student model is trained using a knowledge distillation framework that uses a pre-trained teacher model. An Adam optimizer [4] with an initial learning rate of 0.001 is employed, coupled with a ReduceLROnPlateau scheduler to adapt the learning rate based on loss trends. The batch size is set to 32, and gradient clipping with a max norm of 1.0 is applied to ensure stability. The distillation loss combines cross-entropy on true labels and Kullback-Leibler (KL) divergence between teacher and student logits, weighted by $\alpha = 0.7$. A dynamic temperature scaling mechanism is used, starting at $T = 5.0$ and gradually adjusted over 10 epochs. Training experiments were

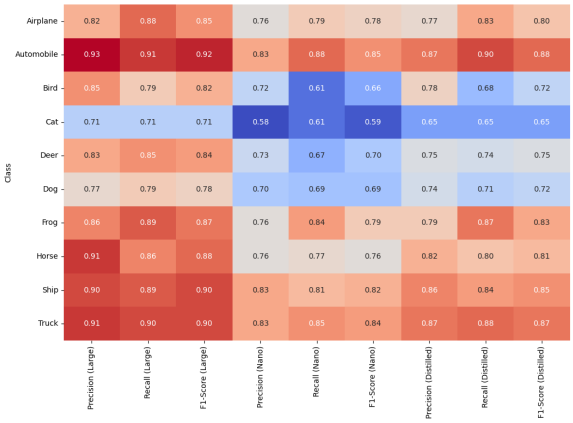

Figure 1. CIFAR-10 Classwise Performance Matrix

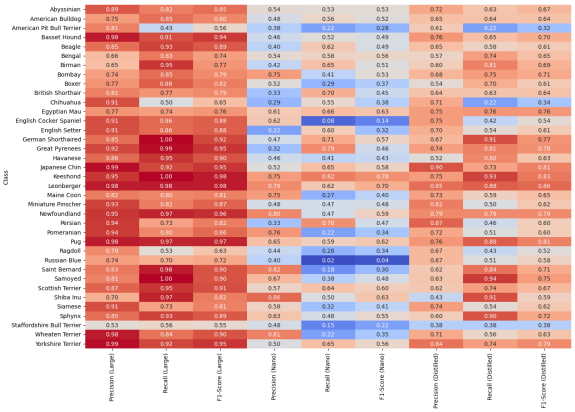

Figure 2. Oxford-IIIT Pets Classwise Performance Matrix

conducted on an NVIDIA 3050Ti laptop GPU using the PyTorch framework.

4.2. Results

Hyperparameters such as the distillation loss weight α, temperature T, and learning rate were optimized through a grid search.

The experimental results reveal clear trends in the effectiveness of knowledge distillation across different datasets, as summarized in Table 1. Below is an analysis of the key findings:

CIFAR-10 Performance Analysis. The distilled YOLOv8Nano model improved accuracy by **5.22%** over its baseline. Notably, this improvement comes with a reduction in inference time from **7.9 ms** to **5.7 ms**. This suggests that knowledge distillation not only enhances the student model's accuracy but also optimizes its computational efficiency. Figure 1 shows a class-wise heatmap of CIFAR-10 performance, illustrating that the model's ability to generalize improved across various classes. Classes with previously lower baseline accuracies, such as *"cat"* and *"horse,"* exhibited noticeable gains, suggesting the transfer of nuanced feature representations from the teacher model.

Tiny-ImageNet Performance Analysis. Tiny-ImageNet presented a more complex challenge, with a **35.36%** relative accuracy improvement in the distilled YOLOv8Nano. The jump from **33.80%** to **43.56%** indicates that the distillation framework effectively bridges the knowledge gap between the teacher and student models for fine-grained tasks. This performance boost, coupled with a reduced inference time, highlights the potential of this approach for real-time, high-complexity visual tasks in resource-constrained environments.

Trade-offs and Efficiency. One of the most compelling aspects of the distilled models is their efficiency. Across all datasets, inference times remained almost the same as the nano model or lower while accuracy increased. This is particularly valuable for deployment in latency-sensitive environments where both speed and precision are critical. The consistent reduction in inference time demonstrates the dual benefit of knowledge distillation in achieving high accuracy without compromising speed.

Overall, the distillation process successfully balances performance and computational efficiency, positioning distilled YOLOv8Nano models as strong candidates for applications in constrained environments, such as autonomous systems and mobile devices. These results reaffirm the broader potential of knowledge distillation in democratizing deep learning by making high-accuracy models accessible in real-world scenarios without the need for significant hardware resources.

4.3. Fine-Grained Classification

Dataset selection and motivation. The Oxford-IIIT Pets [7] dataset was chosen to evaluate the generalization capability of the distilled models on a fine-grained classification task. Unlike CIFAR-10, which primarily focuses on inter-class differences, this dataset comprises 37 classes of pet breeds, including visually similar cats and dogs. Fine-grained classification tasks such as these pose a significant challenge due to the subtle intra-class variations and shared visual features across classes. Notably, the poor performance of the baseline models on CIFAR-10's cat and dog classes, as illustrated in Figure 1, highlighted the need to assess whether knowledge distillation could improve model performance on datasets with subtle distinctions. This transition to a more complex dataset ensures a rigorous evaluation of the student model's ability to generalize beyond coarse-grained tasks.

Relevance of fine-grained evaluation. Generalization to fine-grained datasets is a critical test for lightweight models, particularly in applications where distinguishing subtle visual cues is essential, such as medical imaging or species recognition. Unlike simpler tasks, fine-grained classifica-

Dataset	Model	Params (M)	FLOPs (B)	Accuracy (%)	Inference Time (ms)
CIFAR-10	YOLOv8l Baseline	35.7	99.7	84.67	13.5
	YOLOv8n Baseline	2.7	4.3	75.04	7.9
	YOLOv8n Distilled	2.7	4.3	**78.96**	**5.7**
Tiny-ImageNet	YOLOv8l Baseline	35.7	99.7	57.38	8.6
	YOLOv8n Baseline	2.7	4.3	33.80	4.6
	YOLOv8n Distilled	2.7	4.3	**43.56**	**5.3**
Oxford Pets	YOLOv8l Baseline	35.7	99.7	84.46	13.8
	YOLOv8n Baseline	2.7	4.3	47.62	6.5
	YOLOv8n Distilled	2.7	4.3	**65.49**	**5.3**

Table 1. Performance comparison of YOLOv8l Baseline, YOLOv8n Baseline, and YOLOv8n Distilled on CIFAR-10, Tiny-ImageNet, and Oxford-IIIT Pets datasets.

tion requires the model to learn high-resolution feature representations to effectively separate visually similar classes. By evaluating the distilled YOLOv8Nano on Oxford-IIIT Pets, we aimed to investigate whether the knowledge transfer process enabled the student model to address its observed limitations on CIFAR-10 and extend its capabilities to handle more challenging scenarios. This step also provides insight into the true level of detail that the limited YOLOv8n framework could capture from the teacher model.

Results and observations. The results in Table 1 demonstrate the effectiveness of knowledge distillation for fine-grained classification. The baseline YOLOv8Nano model achieved an accuracy of only 47.62% on the Oxford-IIIT Pets dataset, reflecting its difficulty in learning fine-grained features. In contrast, the distilled YOLOv8Nano model achieved a significant improvement, with an accuracy of 65.49%, representing a 37.61% relative gain. This improvement was achieved without additional computational overhead, as evidenced by the consistent parameter count and reduced inference time of 5.3 ms.

5. Analysis

Challenges of Knowledge Distillation on Smaller Models.

While knowledge distillation has shown considerable improvements in model performance, there are inherent limits to what a smaller model, such as YOLOv8Nano, can achieve due to its reduced computational capacity.

The baseline YOLOv8Nano model exhibits notable performance deficiencies, particularly in handling complex or fine-grained classification tasks. This is reflected in the precision, recall, and F1-score performance metrics, where the distilled model, although improved, still lags behind the large teacher model (YOLOv8Large) across multiple classes. For example, while YOLOv8Large achieves an F1-score of 0.85 for classifying dogs, the distilled model only reaches an F1-score of 0.67, despite showing improvements from the baseline (0.56).

The performance heatmap for both models (as seen in the precision, recall, and F1-score values) reveals a clear disparity, indicating that although knowledge distillation aids in transferring knowledge, the smaller model's limited computational resources restrict its ability to fully capture the subtle distinctions required for accurate fine-grained classification.

Computational Limits and Fine-Grained Classification.

The smaller model, despite its improved performance through distillation, still faces significant challenges in classifying breeds that exhibit minimal intra-class variance. For instance, in classes where fur patterns, shapes, and sizes are similar across breeds, the YOLOv8Nano model struggles to differentiate these subtle features.

The precision for breeds like "Siamese" and "Bengal" cats in the distilled model (0.72 and 0.75 respectively) remains lower than the teacher model's precision values (0.89 and 0.81), suggesting that the distilled model still lacks the fine-grained feature extraction capability of the larger model.

These challenges stem from the fundamental limitations of smaller architectures, which, even with the knowledge distillation process, cannot fully replicate the capacity of larger models to discern such intricate differences. This is a key takeaway when evaluating the effectiveness of knowledge distillation: while distillation significantly boosts performance, it cannot fully overcome the computational and representational constraints of smaller models, especially in highly nuanced tasks.

Generalization Limits. The performance improve-

ments achieved through knowledge distillation are indicative of the potential for generalization, but they also highlight the limits of such methods when applied to tasks with extreme intra-class variability. While the distilled model improves overall accuracy and exhibits better performance in several classes, certain breeds with minimal distinguishing features remain challenging. For instance, breeds such as "Persian" and "Ragdoll" cats still present difficulties, as reflected by the relatively low recall and precision scores in the heatmap, where values hover around 0.40–0.50 for these classes. This underscores the importance of the model's capacity to discern subtle features for high-accuracy classification, which the smaller model is unable to achieve even with distillation. The distilled model may show a higher recall rate than the baseline, but it still fails to reach the consistency and robustness seen in the teacher model, emphasizing the need for more sophisticated strategies like contrastive loss or multi-scale feature learning to enhance feature separability in such fine-grained tasks.

Future Directions. To overcome these limitations, future work could explore enhancing the distilled model's capacity to capture fine-grained distinctions through methods that improve its representational power. Approaches such as multi-resolution input images, contrastive learning techniques, or domain-specific feature augmentation could provide the necessary fine-grained features that the smaller model struggles to capture. Additionally, leveraging external knowledge sources, such as textual descriptions of breeds or advanced data augmentation techniques, could further improve the model's robustness in fine-grained classification. These directions aim to address the challenges highlighted in the error analysis, where the model frequently misclassifies breeds with subtle visual differences. While knowledge distillation can provide a boost, the computational limitations of smaller models necessitate further innovation to bridge the gap in fine-grained tasks.

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 3

[2] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. 2

[3] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics yolov8, 2023. 1

[4] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. 3

[5] Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009. 3

[6] Zheng Li, Xiang Li, Lingfeng Yang, Borui Zhao, Renjie Song, Lei Luo, Jun Li, and Jian Yang. Curriculum temperature for knowledge distillation, 2022. 2

[7] Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012. 4

[8] Shangquan Sun, Wenqi Ren, Jingzhi Li, Rui Wang, and Xiaochun Cao. Logit standardization in knowledge distillation. *arXiv preprint arXiv:2403.01427*, 2024. 2, 3

A. Loss Graphs and Confusion Matrices

A.1. Training Losses for Knowledge Distillation

A.1.1 Oxford-IIIT Pets Dataset Training Loss

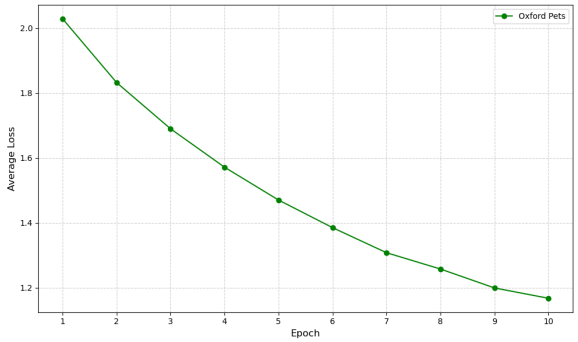

Figure 3. Training Loss curve for the Oxford-IIIT Pets Dataset.

A.1.2 CIFAR-10 Pets Dataset Training Loss .

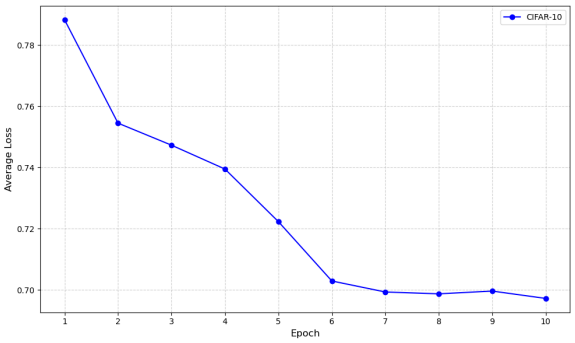

Figure 4. Training Loss curve for the CIFAR-10 Dataset.

A.1.3 Tiny-ImageNet Dataset Training Loss

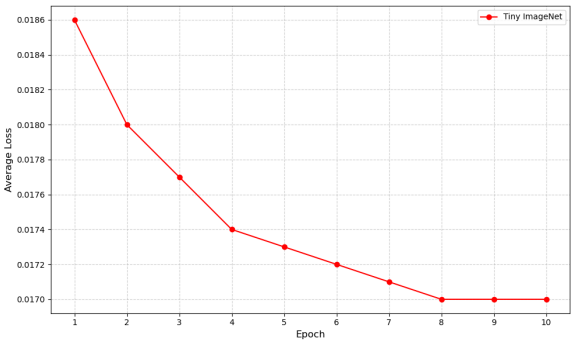

Figure 5. Training Loss curve for the Tiny Dataset.

A.2. Confusion Matrices

A.2.1 CIFAR Dataset

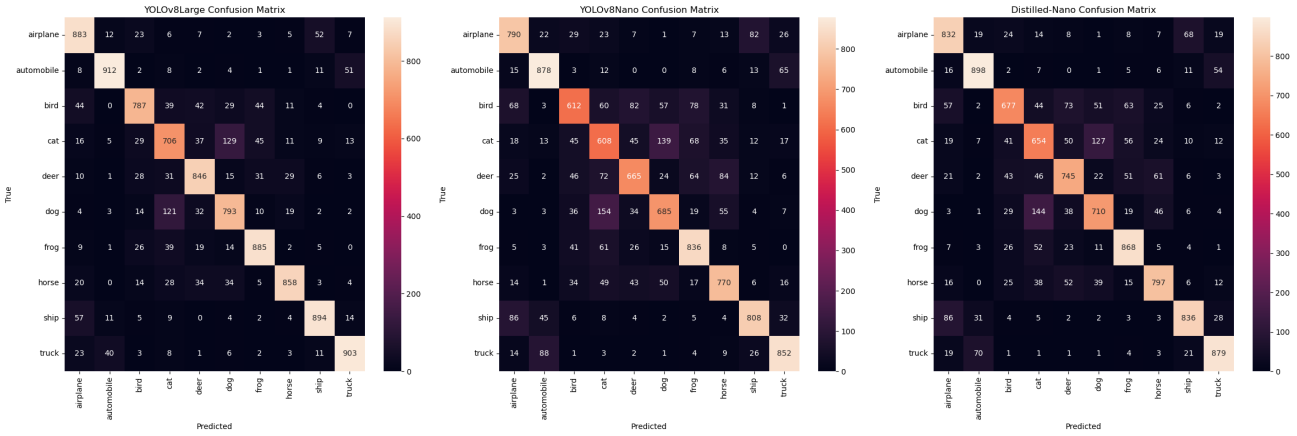

Figure 6. Confusion Matrix for CIFAR Dataset using different models.

A.2.2 Oxford-IIIT Pets Dataset

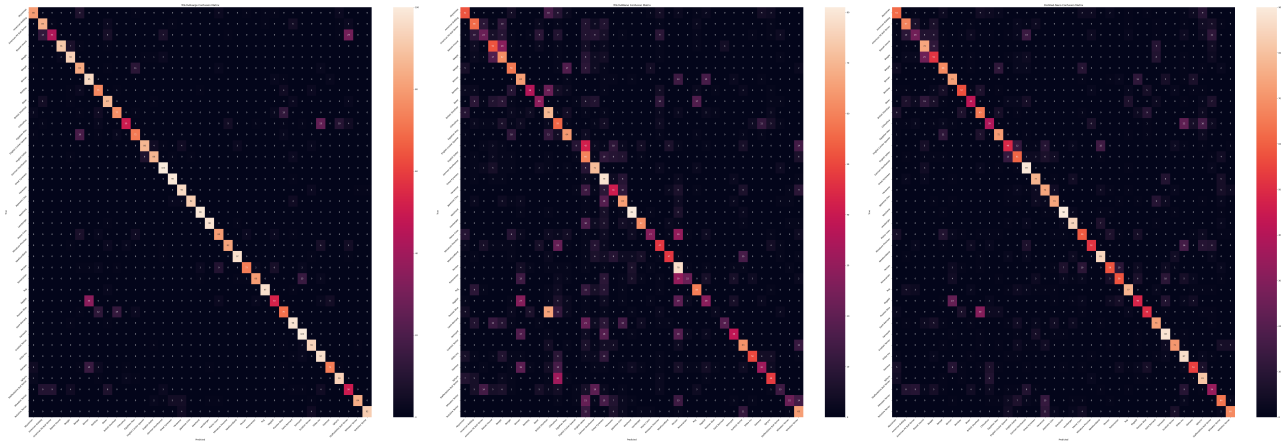

Figure 7. Confusion Matrix for Oxford-IIIT Pets Dataset using different models.