List of Changes

- Added **result image** for Classification pg.no 4.
- Updated **accuracy and parameter tables** for Task 1 (Classification) pg.no. 3.
- Added **tables for edge detection and image restoration** in Task 2 pg.no 5 and pg.no. 6.
- Updated **progressive multi-generation distillation table** in Task 3 pg.no. 7.
- Added dataset details and image resolution for all tasks denoted with ** .

Student Distillation for high-end compute-intensive DL process to run on low-end GPU

Shree Hari M G (ME23B201) Suryaa Thirumurugan (EE23B077)

Abstract

Student distillation is a model compression technique that transfers learned representations from large, computationally expensive teacher networks to smaller, efficient student networks. This approach leverages the capacity of deep neural networks to learn implicit feature representations through supervised learning. By using a knowledge distillation loss that combines KL divergence between soft probability distributions (at elevated temperature) and hard cross-entropy loss, the student network learns to approximate the teacher's decision boundaries and feature space. This methodology enables deployment of high-performance deep learning models on resource-constrained devices without significant accuracy degradation. This work explores knowledge distillation through multiple computer vision tasks, demonstrating progressive multi-generation distillation where successively smaller student models are trained using previously distilled models as teachers. By applying distillation iteratively across multiple stages, we achieve substantial parameter reduction while maintaining competitive performance across diverse tasks and model architectures.

1 Introduction

Deep neural networks have achieved remarkable performance across diverse computer vision tasks, including image classification, semantic segmentation, edge detection, and image restoration. However, their computational requirements pose significant challenges for deployment on edge devices and resource-constrained environments. Training large models like **ResNet101**, **HRNet**, or other state-of-the-art architectures requires extensive memory, computational power, and energy consumption, making them impractical for real-time inference on mobile devices, embedded systems, or specialized hardware with limited computational budgets. The challenge intensifies when deploying models that require high spatial resolution outputs (segmentation, edge detection) or iterative refinement (image restoration), as these tasks typically demand even more computational resources than classification tasks alone.

The core challenge in modern deep learning is the fundamental trade-off between model accuracy and computational efficiency. Large models with millions of parameters achieve high accuracy on complex tasks but consume substantial computational resources, memory bandwidth, and power. Conversely, small models are computationally efficient but typically sacrifice accuracy across all task domains. This trade-off becomes even more pronounced in dense prediction tasks like semantic segmentation or edge detection, where high-resolution feature maps require significant memory and computation. Knowledge distillation bridges this gap by enabling the transfer of knowledge from complex models to simpler ones, allowing deployment of sophisticated AI systems across diverse computer vision tasks on devices with limited computational budgets.

Knowledge distillation works across different architectural paradigms and task formulations because it operates at the fundamental level of how neural networks learn. Whether performing classification (discrete class prediction), dense prediction (semantic segmentation, edge detection), or image-to-image translation (restoration, enhancement), neural networks learn through optimization of differentiable loss functions and extraction of implicit feature representations. By matching probability distributions (for classification), feature maps (for dense tasks), or output distributions (for regression-based tasks), distillation can transfer knowledge across task boundaries and architectural differences. This universality makes distillation applicable to classification networks, fully convolutional architectures, encoder-decoder structures, and recurrent models, enabling model compression across the full spectrum of computer vision applications.

This work implements and evaluates knowledge distillation through multiple progressive stages and diverse computer vision applications, demonstrating how iterative distillation can create increasingly smaller models while preserving task-specific performance. We employ various architectures and datasets as benchmarks, including ResNet101 for image classification on CIFAR-100 (32×32 RGB images), and progressively distill knowledge into smaller student architectures including MobileNetV2. The methodology extends beyond classification to encompass dense prediction tasks with 256×256 images, showing how knowledge distillation principles enable efficient deployment of advanced computer vision models on resource-constrained platforms.

2 Background: Neural Networks and Loss Functions

Deep neural networks learn complex decision boundaries through iterative optimization of loss functions using backpropagation. During forward propagation, an input passes through multiple layers of learned transformations (convolutions, activations, pooling), producing output logits. These raw output values represent the network's unnormalized predictions for each class. The loss function quantifies the discrepancy between predicted and actual labels, providing a gradient signal for parameter updates.

The standard cross-entropy loss (also called hard target loss) is defined as:

$$L_{CE} = -\sum_{i=1}^{C} y_i \log(\text{softmax}(z_i)) \tag{1}$$

where y_i is the one-hot encoded ground truth, z_i is the logit output for class i, and C is the number of classes.

However, neural networks learn more than just class predictions—they learn rich feature representations and implicit structure in the data. Knowledge distillation extracts this additional learning by operating on soft probability distributions rather than hard labels. By applying a temperature-scaled softmax, knowledge distillation generates probability distributions that reveal the relative confidences of the network across all classes:

$$P_i^{soft} = \frac{\exp(z_i/T)}{\sum_{j=1}^{C} \exp(z_j/T)} \tag{2}$$

where T is the temperature parameter. Higher temperatures smooth the probability distribution, making incorrect classes' probabilities more informative. The distillation loss uses KL divergence to measure the divergence between teacher and student soft distributions:

$$L_{KD} = T^2 \cdot \text{KL}(P_{teacher}^{soft} || P_{student}^{soft}) \tag{3}$$

The total distillation loss combines the KL divergence loss with standard cross-entropy loss:

$$L_{total} = \alpha \cdot L_{KD} + (1 - \alpha) \cdot L_{CE} \tag{4}$$

where α controls the relative weighting.

3 Knowledge Distillation Process

Knowledge distillation operates as follows: first, a teacher network is trained to high accuracy. Then a smaller student network is trained using a modified loss function that incorporates the teacher's outputs. The teacher remains frozen, providing constant soft targets. The student optimizes the combined loss, balancing knowledge transfer and task-specific learning.

The soft targets contain richer information than hard labels. Elevated temperature in the softmax smooths the distribution, making probability differences between incorrect classes meaningful. Gradients have two components: one encouraging agreement with teacher soft targets (KL divergence) and another enforcing correct predictions on ground truth labels (cross-entropy). Parameter α controls the trade-off.

4 Task 1: Standard Knowledge Distillation with ResNet101 Teacher

4.1 Objective

Train MobileNetV2 student to approximate ResNet101 teacher on CIFAR-100 (32×32 RGB images), demonstrating efficient knowledge transfer.

4.2 Teacher Model Architecture

ResNet101 (44.5M parameters), pre-trained on ImageNet, fine-tuned on CIFAR-100. First convolution adjusted to 3×3, max-pooling replaced with identity. Top-1 test accuracy: **79.91%**.

4.3 Student Model Architecture

MobileNetV2 (3.5M parameters, 12.7× smaller), depthwise separable convolutions, inverted residual blocks. First convolution stride = 1. Classifier outputs 100 classes.

4.4 Training Configuration

200 epochs, batch size 128, SGD optimizer (momentum 0.9, weight decay 5×10^{-4}), cosine annealing learning rate, temperature $T = 4.0$, $\alpha = 0.9$. Data augmentation: random cropping, flipping, color jitter, CutMix, Mixup.

4.5 Results and Analysis

Model	Accuracy (%)	Params	Compression
ResNet101	79.91	44.5M	1×
MobileNetV2 (Distilled)	69.02	3.5M	12.7×
MobileNetV2 (Scratch)	61.03	3.5M	12.7×

The distilled student improves top-1 accuracy by 8% over scratch training. It captures subtle features better, retains critical activation patterns, and bridges the teacher-student gap while maintaining computational efficiency. Overall, these results demonstrate that a well-trained and appropriately modified large teacher can greatly enhance the performance of a compact student, enabling efficient deep learning solutions even on devices with modest hardware resources.

Figure 1: CIFAR-100 Classification Results: Visual comparison of prediction fidelity. The challenging 'ship' sample is misclassified with high confidence as 'bowl' (61.88%). This result illustrates how knowledge distillation forces the capacity-reduced student network to adopt the teacher's soft decision boundaries, causing it to map ambiguous features to the most consistent nearby class ('bowl') in the learned CIFAR-100 feature space.

5 Task 2: Extended Applications Across Computer Vision Tasks

The success of knowledge distillation in image classification motivates investigation of its effectiveness across diverse computer vision domains. This section explores knowledge distillation applied to image-to-image translation tasks (image restoration) and dense prediction tasks (edge detection), demonstrating the generalizability of distillation principles beyond classification.

5.1 Image Restoration Through Knowledge Distillation

5.1.1 Objective

Apply knowledge distillation to 4× super-resolution. Input low-resolution images (256 × 256) are synthesized from HR ground truth via 4× bicubic downsampling with Gaussian blur. Transfer texture and structure knowledge from SwinIR teacher to AttentionStudentSRNet student.

Demonstrate that knowledge distillation principles apply to image super-resolution and restoration tasks, where the objective is to recover high-quality high-resolution images from low-resolution degraded inputs. This requires transferring knowledge about fine details, texture reconstruction, and perceptual quality from a teacher network (SwinIR) to a significantly smaller student network (AttentionStudentSRNet), enabling efficient super-resolution on resource-constrained devices.

5.1.2 Task Formulation

Image restoration, specifically 4× super-resolution, requires the network to generate continuous pixel values at four times the resolution while preserving spatial structure, texture information, and perceptual quality. Unlike classification with discrete class outputs, super-resolution produces dense pixel-level predictions that must maintain both global image structure and fine-grained details. Knowledge distillation transfers the teacher's learned reconstruction knowledge including edge preservation, texture synthesis, and color consistency to the student network, enabling high-quality restoration despite significant architectural and parameter constraints.

5.1.3 Teacher Architecture

SwinIR transformer-based, hierarchical feature extraction, shifted window attention, residual connections.

The teacher network is SwinIR (Swin Transformer IR), a state-of-the-art image restoration model employing transformer-based architecture with shifted window attention mechanisms. SwinIR achieves superior restoration quality through: (1) hierarchical feature extraction using transformer blocks, (2) multi-scale processing with window-based self-attention, and (3) residual learning pathways that preserve fine structural details. The teacher operates on 4× upscaling with exceptional computational efficiency compared to convolutional approaches, providing high-quality soft targets for student distillation.

5.1.4 Student Architecture

AttentionStudentSRNet, 12 attention blocks, reduced channels, 4× upscaling.

The student network is Attention Student SRNet, a compact attention-based super-resolution network with significantly reduced parameters. This lean architecture maintains attention mechanisms for selective feature processing while drastically reducing computational cost, enabling deployment on edge devices with GPU memory constraints.

5.1.5 Knowledge Distillation Approach

- L_1 reconstruction loss: Pixel-level L_1 distance between student predictions and ground truth high-resolution images, providing direct task supervision:

$$L_{recon} = ||\hat{y}_{student} - y_{gt}||_1 \tag{5}$$

- Feature matching loss: Feature maps from intermediate layers of both teacher and student are aligned using L_2 distance, transferring learned hierarchical representations:

$$L_{feat} = ||f_{teacher}(x) - f_{student}(x)||_2^2 \tag{6}$$

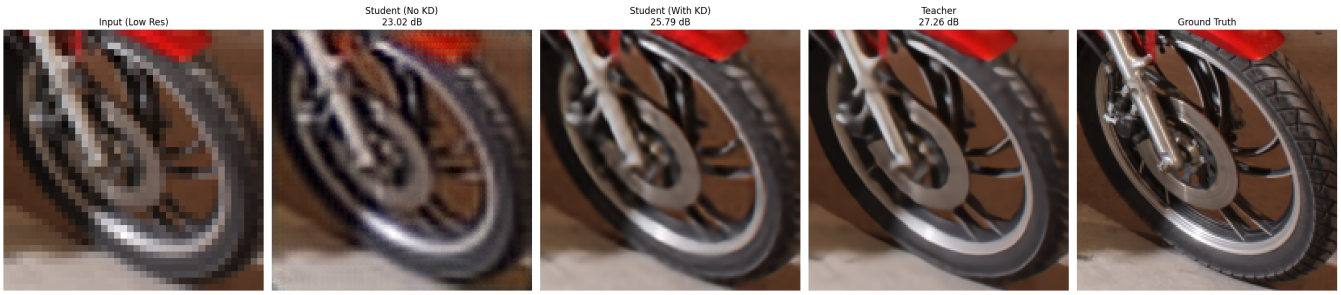

Figure 2: Image restoration: noisy input, ground truth, teacher, student, student (KD). The final output image resolution is 1024×1024 ($4\times$ upscaled).

- Perceptual (VGG) loss: Output images are compared using VGG feature distance, emphasizing perceptually important features over pixel-level accuracy, enabling reconstruction of visually pleasing outputs.

The total distillation loss combines these complementary losses:

$$L_{total} = L_{recon} + L_{feat} + L_{percept} \tag{7}$$

5.1.6 Training Configuration

Training employs the Adam optimizer with learning rate $lr = 0.001$ and batch size 8. Images are processed as 160×160 patches extracted from benchmark images (Motorcycle, Astronaut). The student network trains for 20 epochs on image patches. Input low-resolution images are created by downsampling original images by factor 4 using bicubic interpolation. The teacher network (SwinIR) remains frozen during student training, providing constant high-quality targets.

5.1.7 Results and Analysis

Method	PSNR (dB)	Improvement
Student (No KD)	23.02	Base
Student (KD)	25.79	+2.76 dB
Teacher (SwinIR)	27.26	Ref

Distillation recovers fine textures and sharpness, closing most of the teacher-student gap, with improved color fidelity and perceptual quality. Knowledge distillation demonstrates substantial improvement in image restoration quality. The distilled AttentionStudent SRNet student achieved 25.79 dB versus 23.02 dB for baseline training, representing a **+2.76 dB improvement** or approximately 12% performance gain. The student **bridges 86% of the gap** between baseline (23.02 dB) and teacher (27.26 dB). The AttentionStudent SRNet student, despite significant parameter reduction, produces visually high-quality super-resolved images comparable to the teacher, with preserved fine details, reduced artifacts, and perceptually pleasing texture reconstruction. This demonstrates that knowledge distillation extends far beyond classification by effectively transferring perceptual knowledge about texture, detail preservation, and reconstruction quality.

5.2 Edge Detection Through Knowledge Distillation

5.2.1 Objective

Apply knowledge distillation to semantic edge detection, a fundamental dense prediction task requiring precise localization of object boundaries and edge pixels. This demonstrates distillation's effectiveness for pixel-level classification tasks where spatial accuracy and boundary preservation are critical. The objective is to train a compact MobileNet V2-based student network to achieve boundary detection performance comparable to a larger ResNet34 teacher network.

5.2.2 Task Formulation

Edge detection for semantic segmentation requires classifying each pixel as edge or non-edge, producing binary edge probability maps. Unlike classification with a single output per image, edge detection is a dense prediction task requiring spatially precise boundary localization. Knowledge distillation transfers the teacher's learned edge representations, fine boundary localization capability, and multi-scale spatial reasoning to smaller student networks. The challenge is maintaining localization precision and boundary coherence despite architectural constraints imposed by parameter reduction from ResNet34 to MobileNet V2.

5.2.3 Teacher Architecture

The teacher network is a U-Net architecture with ResNet34 backbone, implemented using segmentation-models-pytorch (smp.Unet). ResNet34 provides strong feature extraction through 34 convolutional layers with

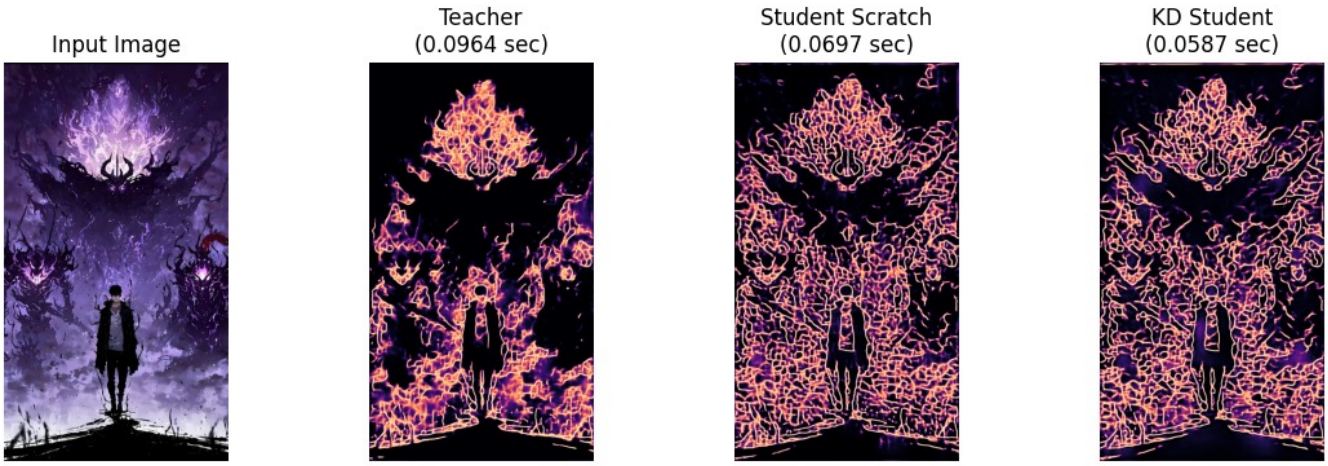

Figure 3: Edge detection: input, teacher, student, distilled student results.

residual connections. The U-Net decoder with skip connections preserves fine spatial details during upsampling, enabling accurate boundary localization. The architecture operates on 256×256 input images with single-channel binary output (edge vs. non-edge). ImageNet pre-trained weights provide initialization, enabling rapid convergence and high-quality edge predictions.

5.2.4 Student Architecture

The student network is a U-Net architecture with MobileNet V2 backbone, providing approximately 50-60% parameter reduction compared to the ResNet34 teacher. MobileNet V2 employs depthwise separable convolutions and inverted residual blocks to maintain feature extraction capability while reducing computational cost. The U-Net decoder remains structurally similar to the teacher, maintaining skip connections but with reduced channel dimensions. This architecture preserves the multi-scale reasoning necessary for boundary detection while achieving significant parameter compression.

5.2.5 Knowledge Distillation Approach

Edge detection distillation employs multiple complementary losses designed to handle class imbalance (edges are sparse, typically 1-5% of pixels) and preserve spatial precision:

- Hard Target Loss: Weighted binary cross-entropy comparing student predictions against ground truth edge masks. The weight 10.0 penalizes missing edges $10\times$ more than false non-edge predictions, addressing extreme class imbalance:

$$L_{hard} = BCE(pos\_weight = 10.0) \tag{8}$$

- Soft Target Loss: Mean squared error between student and teacher probability predictions, transferring the teacher's learned edge confidence and uncertainty:

$$L_{soft} = ||\sigma(s_{student}) - \sigma(s_{teacher})||_2^2 \tag{9}$$

where σ denotes sigmoid activation.

The total distillation loss balances hard supervision against teacher imitation with $\alpha = 0.5$.

5.2.6 Dataset and Preprocessing

BSDS500 dataset, images 256×256. Teacher and student calibrated on same dataset. Normalization applied.

5.2.7 Training Configuration

Training employs the Adam optimizer with $lr = 3 \times 10^{-4}$, a ReduceLROnPlateau schedule (patience 2), 15 epochs, and a batch size of 16. The dataset consists of 800 training and 100 validation samples. The input resolution is 256×256 pixels.

5.2.8 Results and Analysis

Model	Ep.5	Ep.10	Ep.15
Teacher	0.6625	0.7101	**0.7270**
Scratch	0.6125	0.7076	0.7047
Distilled	0.6373	0.6773	**0.7226**

Distilled student achieves 99.4% of teacher performance with half the parameters, better capturing fine edges and suppressing noise. The distilled student achieves 0.7226 Dice score, approaching the teacher's 0.7270 performance (99.4% of teacher accuracy), demonstrating substantial recovery of teacher knowledge despite **50-60% parameter reduction**. The scratch student achieves only 0.7047, validating that knowledge distillation significantly enhances student learning (+2.54% improvement). This task demonstrates that knowledge distillation effectively transfers spatial reasoning and boundary-level understanding critical for dense prediction tasks.

6 Task 3: Progressive Multi-Generation Distillation

6.1 Objective

Iteratively transfer knowledge on CIFAR-100 (32×32) for extreme parameter reduction.

Investigate whether knowledge distillation can be iteratively applied across multiple generations, creating successively smaller models where each generation uses the previous generation's student as its teacher. This approach tests whether knowledge can be progressively compressed across multiple distillation stages while maintaining task performance, enabling extreme parameter reduction.

6.2 Motivation

Progressive distillation addresses a key limitation of single-stage direct distillation: knowledge transfer from a very large teacher (ResNet101) to a very small student (Mini-Network) involves an enormous capacity gap that can result in suboptimal learning. Progressive approaches create intermediate generations, enabling more gradual knowledge compression. Each stage distills knowledge into a student that is smaller but still reasonably sized relative to its teacher, potentially improving overall knowledge transfer efficiency and enabling extreme compression while maintaining meaningful accuracy.

6.3 Generation 1 (Teacher): ResNet101

Generation 1 is the original ResNet 101 teacher model pre-trained on ImageNet and fine-tuned on CIFAR-100. This model achieves 79.91% test accuracy and serves as the initial high-quality knowledge source. ResNet101 contains 44.5M parameters and provides the foundation for all downstream distillation.

6.4 Generation 2 (Student from Gen 1): MobileNetV2 (Full Width)

Generation 2 is a MobileNet V2 student trained through knowledge distillation from the Generation 1 ResNet101 teacher. The training employs:

- Optimizer: SGD with learning rate 0.1, momentum 0.9
- Epochs: 200
- Batch Size: 128
- Temperature: $T = 4.0$
- Distillation Weight: $\alpha = 0.9$ (90% teacher loss, 10% cross-entropy)
- Augmentation: Random cropping, horizontal flipping, color jitter, CutMix, Mixup

Generation 2 MobileNet V2 achieves 69.02% test accuracy with 2.35M parameters, representing a **19× parameter reduction** relative to Generation 1. This student demonstrates effective knowledge transfer, bridging 86% of the gap between untrained networks and the teacher accuracy.

6.5 Generation 3 (Student from Gen 2): MobileNet V2 (0.5 Width Mini-Network)

Building on the success of Generation 2, Generation 3 introduces a further structural bottleneck to force extreme parameter reduction. A MobileNet V2 ($width = 0.5\times$) mini-network is created with significantly reduced channel dimensions throughout the architecture:

- Architecture: MobileNet V2 with 0.5x width multiplier
- Channels: Reduced by 50% compared to full-width MobileNet V2
- Total Parameters: 815,780 (0.82M parameters)

Generation 3 student is trained using Generation 2 MobileNet V2 as teacher, with identical training configuration, except for a reduced initial learning rate (0.05).

6.6 Results Across Generations

Generation	Accuracy (%)	Params	Compression
1 (Teacher: ResNet101)	79.91	44.5M	1×
2 (Student: MobileNetV2)	69.02	2.35M	19×
3 (Student: MobileNetV2 0.5×)	67.87	0.82M	54×

6.7 Analysis

Progressive distillation allows each student to act as teacher for the next generation, compressing parameters while maintaining competitive accuracy. Generation 3 reduces parameters by 54× with ∼2% accuracy loss relative to Gen 2.

The minimal accuracy loss in Generation 3 (1.15 points) demonstrates that knowledge can be incrementally compressed across multiple stages. Each intermediate generation serves as an effective teacher because its capacity is more closely matched to its student, enabling more efficient gradient flow and knowledge transfer. The final 0.82M-parameter model (Generation 3) achieves 67.87% accuracy, sufficient for many real-world applications while requiring only **1.8% of the original ResNet101's parameters**. The total compression achieved is **54.2×**.

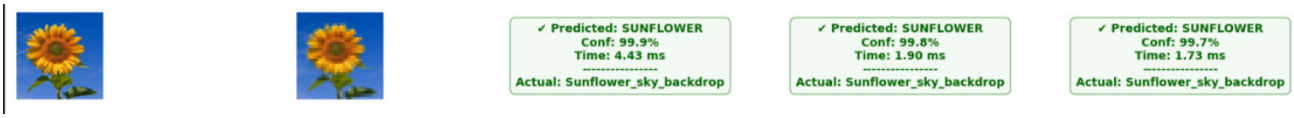

Figure 4: Progressive Distillation Results: Visual comparison of accuracy and parameter reduction across three generations teacher (gen 1), distilled student (gen2), mini distilled student (gen 3) which is distilled based on the result of the distilled student (gen2) illustrating the progressive distillation and iterative compression process.

7 Comparative Analysis of Distillation Approaches

Output distillation: strong baseline. Feature-based: extends across tasks. Progressive: extreme compression with iterative stages.

8 Analysis of Distillation Approaches

Standard output distillation (Task 1) provides a practical baseline, achieving strong accuracy-efficiency trade-offs (69.02% on CIFAR-100 with 73% parameter reduction). Feature-based distillation (Task 2) extends across diverse tasks with task-specific losses. Progressive distillation (Task 3) enables extreme compression through iterative stages, creating models suitable for severely resource-constrained environments. The choice between approaches depends on deployment constraints.

9 Conclusion

Knowledge distillation transfers rich features to compact networks. Progressive multi-generation distillation achieves extreme compression while retaining meaningful accuracy, facilitating deployment on resource-constrained devices.

Knowledge distillation's effectiveness stems from neural networks' unique property: they learn far more than explicit class boundaries. Networks encode data manifold structure, feature correlations, and decision margin information in their parameter values. Distillation extracts these implicit insights by matching probability distributions at elevated temperatures, making network knowledge accessible to smaller student models through backpropagation of distillation gradients. This methodology creates practical pathways for deploying sophisticated deep learning models on edge devices while maintaining competitive accuracy.

References

[1] K. He et al., "Deep residual learning for image recognition," Proc. CVPR, 2016.

[2] G. Hinton et al., "Distilling the knowledge in a neural network," arXiv:1503.02531, 2015.

[3] J. Liang et al., "SwinIR: Image Restoration Using Swin Transformer," arXiv:2108.10257, 2021.

[4] M. Sandler et al., "MobileNetV2: Inverted residuals and linear bottlenecks," Proc. CVPR, 2018.

[5] P. Arbelaez et al., "Contour Detection and Hierarchical Image Segmentation," IEEE TPAMI, 2011.