Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2024.0429000

Robust Knowledge Distillation for Resource-Constrained Applications: Evaluating Offline KD Techniques for Large–Compact Heterogeneous CNNs

CHINMAYA KAUNDANYA[1], PAULO CESAR[2], BARRY CRONIN[2], ANDREW FLEURY[2], MINGMING LIU[1] AND SUZANNE LITTLE[1]

[1]Research Ireland Insight Centre for Data Analytics, Dublin City University, Dublin, D09 V209, Ireland
[2]Luna Systems, Dublin, D11 KXN4, Ireland

Corresponding author: Chinmaya Kaundanya (e-mail: chinmaya.kaundanya3@mail.dcu.ie).

This research was conducted with the financial support of Research Ireland (Grant No. 12/RC/2289_P2), at the Research Ireland Insight Centre for Data Analytics at Dublin City University, and Luna Systems.

ABSTRACT Deploying deep learning models on severely resource-constrained edge platforms, such as microcontroller-based systems, remains a significant challenge due to strict limitations on memory, computation, latency, and power consumption. These constraints necessitate the use of ultra-compact neural networks, which typically suffer from reduced representational capacity and degraded performance compared to standard large-scale models. Knowledge distillation offers a promising solution by transferring information from a high-capacity teacher model to a compact student, improving not only accuracy but also robustness, which is critical for real-world tiny machine learning (TinyML) applications operating under distributional shifts. However, when student models are extremely compact, the large capacity gap between teacher and student can hinder effective knowledge transfer. In standard distillation settings, this gap can be narrowed by simply using a larger student. Resource-constrained deployments, however, impose strict upper bounds on model size, making large capacity gaps inherent and unavoidable. In this work, we systematically investigate the effectiveness of three main categories of offline knowledge distillation including response-based, feature-based, and relation-based across heterogeneous teacher–student convolutional neural network (CNN) pairs. We specifically target ultra-compact student architectures representative of TinyML deployments, examining how varying capacity gaps affect distillation in this setting. We first evaluate classification accuracy on the CIFAR-100 benchmark and then assess robustness on the CIFAR-100-C dataset without additional fine-tuning. The study is then extended to a practical Driver Monitoring System (DMS) application using the State Farm Distracted Driver Detection dataset, followed by robustness evaluation on the unseen 100-Driver dataset. Our results demonstrate that knowledge distillation remains effective even for ultra-compact CNNs designed for resource-constrained platforms. We further show that even when capacity gaps are consistently large, the relative difference between teacher and student still meaningfully influences distillation effectiveness, with smaller gaps yielding larger accuracy improvements. Among the evaluated methods, relation-based distillation, particularly contrastive representation distillation (CRD), consistently provides the strongest gains in both accuracy and robustness, making it well suited for resource-constrained edge deployments.

INDEX TERMS Knowledge Distillation, compact model, resource-constrained application, TinyML

I. INTRODUCTION

Knowledge Distillation (KD) has emerged as a powerful and flexible paradigm for training compact and efficient deep learning models [1]. The core principle involves a "teacher–student" framework, where a smaller, often computationally inexpensive student model is trained to mimic the behavior of a larger, more complex teacher model. Instead of learning solely from ground-truth labels, the student leverages

''dark knowledge'' [2] from the teacher, which can take the form of soft probability distributions (logits) or intermediate feature representations. This approach enables the student to learn richer and more nuanced data distributions and decision boundaries than it could by training on labeled data alone. This enables the student to achieve performance that closely approaches that of its much complex teacher. The applications of KD extend beyond model compression to include improving model robustness [3], facilitating cross-modal transfer learning [4]–[6], and distilling the knowledge of multiple models or ensembles into a single, deployable network [7], [8].

The utility of knowledge distillation becomes especially pronounced in model compression for deployment on resource-constrained devices. Advances in low-power embedded devices and machine learning (ML) model optimization have enabled tiny machine learning (TinyML). TinyML runs ML models directly on IoT devices to reduce latency, enhance privacy, lower costs, and ensure reliable operation even with limited connectivity through on-device analytics [9]. Modern deep learning models, despite their impressive accuracy, often contain hundreds of millions of parameters and require substantial computational resources, memory, and power, making them impractical for on-device inference. KD addresses this limitation by producing lightweight student models that preserve much of the teacher's predictive performance while meeting the tight operational constraints of embedded hardware. Through this process, advanced AI functionalities, such as real-time image classification [10] or object detection [11], speech processing [12], or environmental monitoring, can be efficiently executed at the edge, with minimal latency [13], reduced power consumption, and enhanced data privacy by limiting cloud dependency [14].

The objectives and constraints of knowledge distillation diverge significantly when comparing standard applications to those for resource-constrained devices, though the foundational teacher-student principles remain consistent. In a typical setting, the primary goal of KD is to maximize the accuracy of the student model, with secondary considerations to improve calibration or generalization. This process usually operates with loose constraints on model size or runtime, focusing on transferring the teacher's knowledge through its soft targets or feature representations.

In contrast, KD for constrained environments, such as edge AI and TinyML, represents a multi-objective optimization problem where accuracy is just one of several critical factors. Optimization must also aggressively target hardware-related metrics, including minimizing the memory footprint, reducing inference latency, and lowering energy consumption per inference. Consequently, success is not defined solely by an accurate student model but by one that performs effectively within strict power and computational budgets, requiring a co-design process that explicitly balances predictive performance with these operational constraints to produce models that are both intelligent and deployable. However, such design constraints inherently create a large performance gap between the teacher and the student models, which introduces an additional challenge during knowledge distillation.

This disparity in complexity and representational power between a large, high-capacity teacher model and a much smaller, compact student model is known as the *capacity gap*, which can cause larger teachers to negatively impact distillation when the student's expressive power is not sufficient to capture the teacher's knowledge [15], [16]. This gap poses a substantial problem because the fundamental goal of many distillation methods is to transfer knowledge by forcing the student to mimic the teacher's outputs or intermediate feature representations. When the capacity gap is large, the student model may be too simple architecturally to effectively learn or replicate the complex, high-dimensional functions and nuanced data representations captured by the teacher. This mismatch can cause knowledge transfer to fail as the rich information from the teacher can be perceived as noise by the student, leading to a difficult optimization problem and poor convergence.

In this work, the teacher–student capacity gap is quantified using parameter count ratios, as model size represents a direct and practical constraint for deployment on low-spec devices. While parameter count does not fully characterize representational capacity, it provides a deployment-relevant proxy in resource-constrained settings. Different CNN architectures exhibit varying inductive biases and parameter efficiency, meaning models with similar sizes may still differ in expressive power. Accordingly, parameter ratios are used here as a pragmatic measure of capacity gap, reflecting realistic hardware constraints rather than intrinsic representational capacity.

In resource-constrained settings, particularly where the student must not only match accuracy but also meet severe constraints on memory footprint, latency, and power, the capacity gap becomes especially problematic. The smaller student must learn from a teacher whose complexity it cannot hope to replicate, yet it must still operate within tight hardware and runtime budgets. To mitigate this issue by gradually bridging the gap between teacher and student, some studies [8], [17]–[19] introduced teacher assistant (TA) models with the philosophy of expanding the single distillation task into multiple easier distillation tasks. However, even though multiple studies address the capacity gap, a research gap still remains for large capacity gaps involving ultra-compact CNN student models in resource-constrained applications. In this work, we evaluate different categories of offline knowledge distillation [1] with various capacity gaps between the teacher and student, both being heterogeneous CNN models.

A compelling example of resource-constrained use of machine learning models on embedded systems where such distillation would be beneficial is found in detection and monitoring of driver distraction. In such driver monitoring systems (DMS), based on image classification, models must operate reliably under strict latency, power, memory, and compute constraints while maintaining the accuracy necessary for meaningful safety functionality. The true measure of a driver behavior

monitoring system lies in its ability to generalize across various drivers, varying camera positions, and challenging illumination conditions. In practical deployments, this also requires ensuring the *robustness* of the model. By *robustness*, we refer to the model's ability to maintain decent accuracy despite changes in appearances, environmental conditions, or other real-world variations that naturally occur in driving scenarios, which in this work is specifically examined through robustness to common image corruptions and dataset-level domain shifts.

To explore this, we examine the effectiveness of knowledge distillation as a technique to improve model efficiency and generalization under such constraints. Specifically, we investigate three representative categories of KD – response/logit-based, feature-based, and relation-based – initially evaluating them on the validation set of the widely used CIFAR-100 dataset [20], which offers a diverse set of object classes for benchmarking. Subsequently, we assess the robustness of these KD categories on the highly augmented CIFAR-100-C dataset [21], without any fine-tuning, to analyze their behavior under distributional shifts.

Finally, we extend our evaluation to a real-world application in driver monitoring: we first test on the State Farm Distracted Driver Detection dataset [22], which provides realistic scenarios for real-time distracted driver detection, and further evaluate robustness on the 100-Driver dataset [23], a distinct and unseen dataset widely used for distracted driver detection research. Across both CIFAR-100-C and 100-Driver, robustness is quantitatively assessed by comparing the performance degradation of the student models, evaluated without fine-tuning on these datasets, using the Macro F1 score. While TinyML deployment involves additional hardware considerations such as latency, runtime memory, and energy consumption, this study focuses on analyzing distillation behavior under strict model size constraints. We therefore target ultra-compact student models that satisfy typical MCU memory budgets, and examine how different KD categories affect their accuracy and robustness, rather than performing hardware-level benchmarking.

Based on this motivation, the key contributions and findings of this study are summarized as follows:

- Knowledge distillation remains effective even for highly compact CNN student models intended for severely resource-constrained platforms. Even the smallest models in our experiments show meaningful gains in accuracy after distillation.
- The capacity gap between the teacher and student plays a clear role in how much improvement distillation provides. The largest accuracy gains occur when this gap is relatively small. Fig. 1 provides a visual impression of the capacity gap between the largest teacher network and the smallest student network used in this study.
- Across the techniques evaluated, relation-based methods – especially the Contrasive Representation Distillation (CRD) [24] approach consistently deliver the most notable improvements in both accuracy and robustness for ultra-compact student models in these constrained settings.

The remainder of this paper is organized as follows. Section II reviews previous work on knowledge distillation for resource-constrained applications and identifies the specific research gap this study addresses. Section III describes the methodology for systematically evaluating how different KD categories impact ultra-compact CNN models. Section IV provides the experimental setup, including datasets, model configurations, and implementation details. Section V presents and discusses the results against our defined evaluation criteria. Finally, Section VI summarizes the key findings, acknowledges limitations, and outlines directions for future work.

II. RELATED WORK

This section presents an overview of prior research relevant to our study, organized around three core areas. We begin by reviewing knowledge distillation methods that address the capacity gap between large teacher networks and much smaller student models. We then discuss distillation techniques developed for resource-constrained edge devices, where compact models are essential for efficient deployment. Finally, we review work applying knowledge distillation specifically to distracted-driver monitoring, highlighting how lightweight models are designed for safety-critical, real-time environments.

A. KNOWLEDGE DISTILLATION FOR HIGH CAPACITY GAP BETWEEN TEACHER AND STUDENT

Recent research in knowledge distillation has increasingly focused on addressing the ''capacity gap'' – a problem that arises when a large, high-capacity teacher model transfers knowledge to a much smaller student model [17]. This performance drop occurs because the student, with its limited capacity, struggles to mimic the complex functions and representations learned by the over-parameterized teacher.

A primary approach to mitigating the capacity gap involves modifying the teacher model to make its knowledge more accessible to the student. The AID [25] method addresses this by fine-tuning the pre-trained teacher to produce an intra-class distribution that is better aligned with the student's capacity before the distillation process begins. DFPT-KD [26] replaces the standard distillation process with a novel ''dual-forward path teacher'', an additional prompt-based forward pass that is established within the teacher and optimized to make the transferred knowledge more compatible with the student's representational ability. Park et.al. [27] propose a method to optimize teacher models before standard distillation. During optimization, an auxiliary sub-network attached to the teacher during teacher training defined as ''student branch'' is jointly learned to create representations that are inherently more suitable for knowledge transfer to smaller student models.

Other methods focus on improving the knowledge transfer process itself to better accommodate the capacity difference. Some progressive distillation techniques [8], [17] consist of a series of intermediate-sized models, often called ''teacher

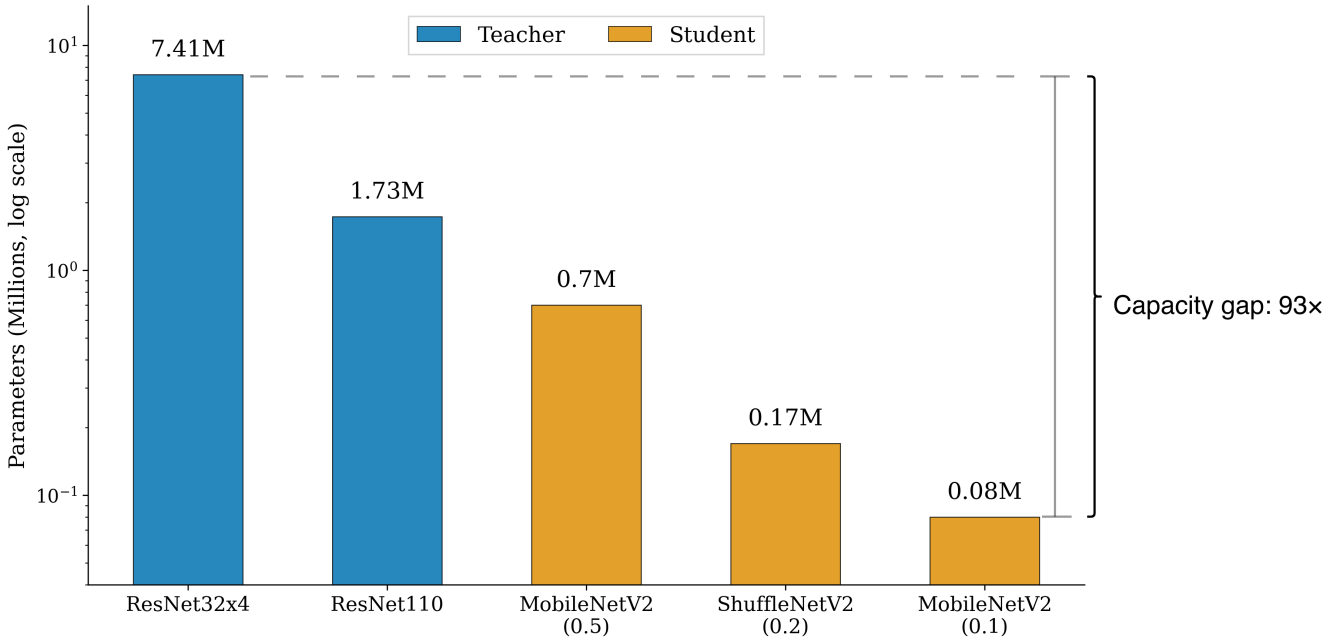

FIGURE 1. Illustration of the capacity gap between the largest teacher and smallest student models used in this study. Considering the 10-class State Farm Distracted Driver Detection dataset, the ResNet32x4 teacher has 7.41M parameters, while the MobileNetV2 (0.1) student has 0.08M parameters, resulting in a capacity gap of 93×.

assistants''. Knowledge is progressively distilled through this chain of models, from the large teacher to the final small student, making the transfer more gradual and effective. ReviewKD [28] extends the standard knowledge distillation pipeline with an additional ''review phase'' for the application of crowd counting. After the teacher transfers its latent features to the student in the instruction phase, the review phase enables the student to refine its initial estimates and improve the predicted crowd density maps using the previously learned features. Identifying the teacher-student confidence gap issue, Jia et. al. [15] propose representing both teacher and student outputs on a common hypersphere. Their spherical knowledge distillation method reduces the harmful impact of confidence magnitude on the student and improves performance.

B. KNOWLEDGE DISTILLATION FOR RESOURCE-CONSTRAINED EDGE DEVICES

Applying knowledge distillation to edge devices introduces specific challenges, primarily due to the limited computational power, memory, and energy resources available. Research in knowledge distillation for this area focuses on creating lightweight, yet powerful models suitable for real-time applications on the edge.

AAKD [29] introduces a strategy 'knowledge sample selection' that dynamically selects the most informative and appropriate training examples for distillation based on the student's current representational ability. It also incorporates an adaptive teacher switching mechanism that automatically transitions between teacher networks as the student grows more capable, enabling more effective knowledge transfer and better performance in compact student models. Khan et.al. [30] propose a weighted knowledge distillation technique for a resource-constrained crowd-counting application using drones. This technique improves the learning capability of lightweight models, allowing them to emulate deeper and more accurate teacher models without the associated computational cost. ATMS-KD [31] uses capacity-aware adaptive temperature scheduling for compact student or larger teacher-student capacity gaps. The process starts with higher 'Temperature' parameter then annealing it over training to gradually shift emphasis from soft to hard labels. Zhao et. al. [32] propose a two-step technique for remote-sensing scene classification on edge devices. First, it distills a large model into a smaller, lighter version via knowledge distillation. Then it adds an early-exit mechanism so that ''easy'' input scenes can be classified fast, exiting the model early, while ''harder'' ones go deeper, so overall it speeds up inference and lowers energy use while maintaining good accuracy.

C. KNOWLEDGE DISTILLATION FOR DISTRACTED DRIVER MONITORING

In the driver-monitoring domain, knowledge distillation has increasingly been leveraged to enable compact models without significant accuracy losses in tasks such as driver behavior and drowsiness detection. For example, Liu et al. [33] use KD in conjunction with neural architecture search (NAS) to guide a student network by first using the teacher to inform architecture search and then distilling the teacher's logits to a student with ∼0.42 M parameters. In the distracted-posture classification setting, Li et al. [34] propose an instance-specific

multi-teacher KD scheme. Unlike uniform teacher weights, their method dynamically assigns teacher weights on a per-instance basis and distills both high- and intermediate-level features to a lightweight CNN for edge deployment.

Tanama et al.'s ''Quantized Distillation'' [35] combines soft-label KD from a large I3D architecture teacher model to a 3D MobileNet student and applies quantization of weights/activations to further shrink and accelerate the model, achieving $\sim 3\times$ size reduction, $\sim 1.4\times$ speed-up. Ahuja et al. [36] uses a large teacher CNN model for drowsiness classification (i.e., open vs closed eyes) and transfers its soft-logit outputs via knowledge distillation to a compact student for real-time inference on edge platforms. Although the work by Gupta et.al. [37] on eye/mouth spatial-temporal features for drowsiness detection does not explicitly emphasize KD, it illustrates the trend towards lightweight feature-based models for drowsiness and could serve as a complementary baseline. Collectively, these works illustrate how KD via soft-label annotations, teacher-student architectures, and in some cases quantization or multi-teacher schemes has become a core methodology for achieving high-accuracy, low-latency models suited for driver-behavior and drowsiness recognition in constrained environments.

This previous work all addresses the common and important aspect of KD – the capacity gap between teacher and student, and demonstrates the application of KD in resource-constrained use-cases such as distracted driver monitoring. However, there remains a clear gap in studying how different degrees of capacity gap (i.e., from a high-capacity teacher to a very compact student heterogeneous CNN models) influence the effectiveness of KD, especially when the student is meant to be deployed on a low-spec edge platform. Our work addresses this gap and additionally investigates the critical factor of how various KD techniques perform in making such highly compact student models robust in real-world, resource-constrained applications characterized by data in diverse distributions with high domain shift.

III. METHOD

In this section, we describe the methodology used to analyze how different knowledge distillation techniques perform in the case of ultra-compact CNN models for resource-constrained applications. Our goal is to understand how various forms of distilled knowledge from larger teacher models influence the accuracy and robustness of compact student models. We begin by explaining the categorization of the offline KD methods used in this study as summarized in Table 1, then outline the common KD setup tailored for resource-constrained hardware, followed by details of the datasets and evaluation strategy.

A. CATEGORIZATION OF OFFLINE KD TECHNIQUES

Offline knowledge distillation methods generally focus on knowledge types and distillation strategies. Knowledge types can be grouped into three categories: response-based, feature-based, and relation-based. Response-based methods transfer knowledge from the teacher's final outputs to guide the student's predictions. Feature-based methods leverage intermediate representations to help the student capture meaningful semantic information across hidden layers. Relation-based methods go beyond individual samples by modeling cross-sample relationships within the dataset. Fig. 2 shows a schematic overview of this categorization.

1) Response-Based KD

Response-based / Logit-based KD distills knowledge from the teacher's output probabilities (logits) by training the student to match the softened predictions, usually via Kullback–Leibler divergence (KL-divergence) [2] with temperature (T) scaling. It is simple and effective, and requires only the teacher's raw outputs. However, it is limited since it captures knowledge only from the probability distributions of final raw logits.

$$L_{ResD}\left(p\left(z_t,T\right),p\left(z_s,T\right)\right)=\mathcal{L}_R\left(p\left(z_t,T\right),p\left(z_s,T\right)\right) \quad (1)$$

Here, $p(z_t,T)$ and $p(z_s,T)$ represent the softened probability distributions obtained from the teacher and student logits, respectively, using temperature-scaled softmax. The loss term $\mathcal{L}_R\big(p(z_t,T),,p(z_s,T)\big)$ is typically implemented using KL divergence, making the student logits, z_s, match the teacher logits, z_t, during distillation [38].

2) Feature-Based KD

Feature-based KD transfers knowledge from the teacher's intermediate layer representations by aligning student feature maps with teacher's corresponding features using similarity losses (e.g., L2 or cosine). This provides richer supervision and captures intermediate abstractions, though it may require projection layers to handle mismatched feature dimensions.

$$L_{\text{FeaD}}\left(f_t(x),f_s(x)\right)=\mathcal{L}_F\left(\Phi_t\left(f_t(x)\right),\Phi_s\left(f_s(x)\right)\right), \quad (2)$$

Here, $f_t(x)$ and $f_s(x)$ denote the intermediate feature maps produced by the teacher and student models, respectively. The transformation functions $\Phi_t(f_t(x))$ and $\Phi_s(f_s(x))$ are applied when the feature maps of the teacher and the student differ in shape ensuring they can be aligned for comparison. The term $\mathcal{L}_F(\cdot)$ represents the similarity function used to match the transformed feature maps during distillation [38].

3) Relation-Based KD

Relation-based KD distills structural knowledge by preserving relationships between samples (e.g., distances, angles, correlations) in the teacher's high dimensional representation space. It captures deeper relational information beyond features or logits but is often more computationally expensive.

$$L_{RelD}\left(F_t,F_s\right)=\mathcal{L}_{R^2}\left(\psi_t\left(t_i,t_j\right),\psi_s\left(s_i,s_j\right)\right), \quad (3)$$

Here, $(t_i,t_j)\in F_t$ and $(s_i,s_j)\in F_s$, where F_t and F_s denote the sets of feature representations produced by the teacher and student models, respectively. The functions $\psi_t(\cdot)$ and $\psi_s(\cdot)$ measure the similarity between the feature pairs (t_i,t_j) and

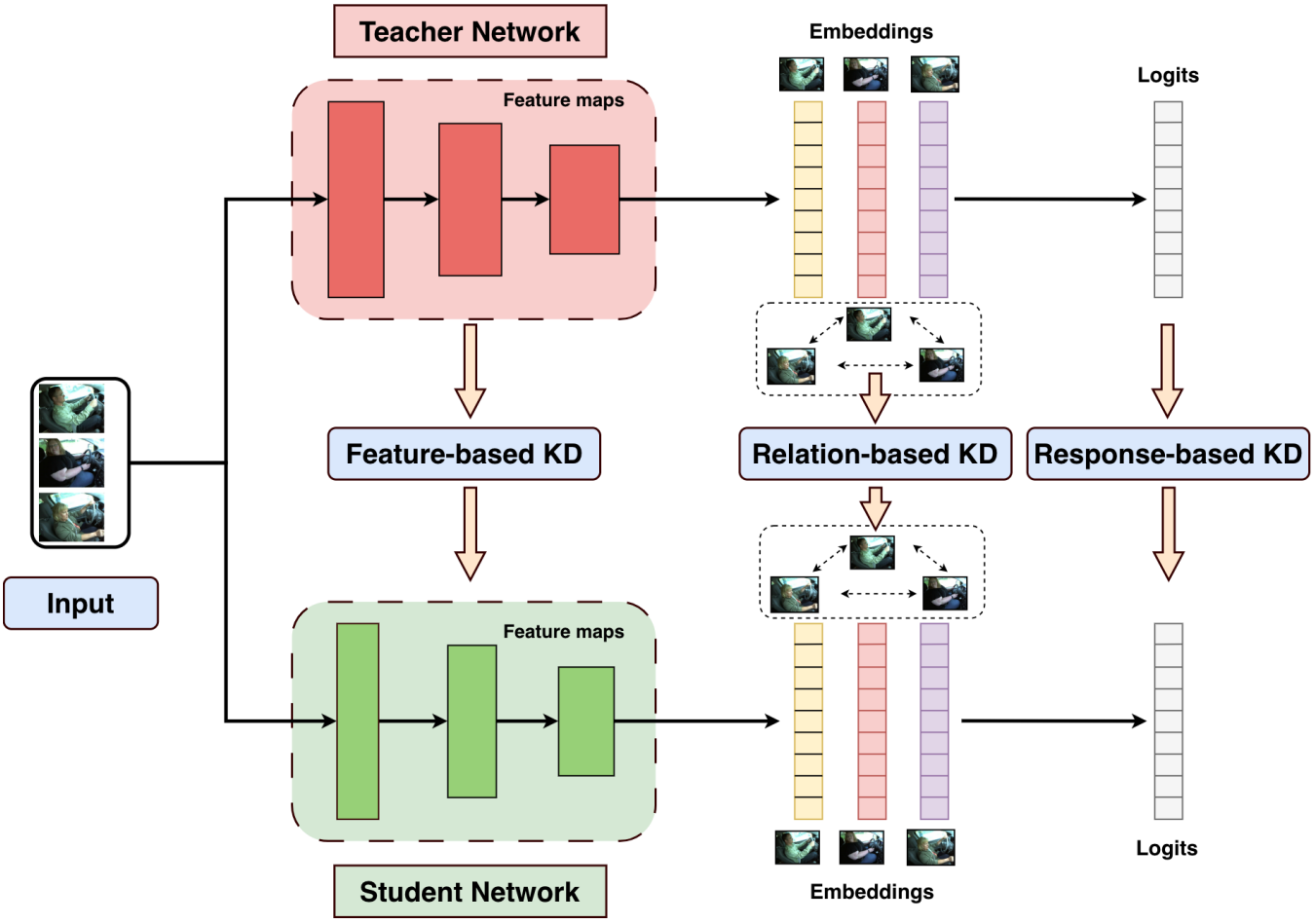

FIGURE 2. A simplified visual overview of knowledge distillation categories: response-based KD transfers knowledge by comparing the teacher and student's final logits, feature-based KD utilizes representations from intermediate feature maps and relation-based KD transfers inter-sample relations in the embedding space.

(s_i, s_j). The term $\mathcal{L}_{R^2}(\cdot)$ represents the correlation-based loss used to align the relational structure of the student with that of the teacher [38].

B. COMMON KD SETUPS FOR RESOURCE-CONSTRAINED APPLICATIONS

1) Highly compact student models

Tiny IoT devices, such as microcontrollers, often provide only ∼256KB of SRAM [48]. In this work, the smallest student model satisfies this lowest-spec constraint demonstrating deployability on such devices. The other student models remain compact (<1MB of parameters, including weights and biases) and are likewise aimed at very low-spec Microcontroller units (MCU), even if not as small as the minimum configuration. Given these extreme memory constraints, this study focuses on compact CNN architectures. CNNs are well-suited to MCU deployment because their core operators, such as convolutions and depthwise separable convolutions, are commonly supported by existing embedded firmware and inference libraries. In contrast, Vision Transformer (ViT) models rely on global self-attention, which introduces higher memory and computational overhead [49]. Recent efforts such as Tiny-ViT aim to improve transformer efficiency for edge devices [50]. However, their resource requirements still exceed the ultra-compact MCU budgets targeted in this work. The accuracy–latency trade-off is a persistent and evolving challenge on resource-constrained platforms. Therefore, our focus is on improving the accuracy and robustness of student models that already meet typical MCU memory requirements.

2) High capacity gap between the teacher and the student

In typical knowledge distillation setups, the teacher model is a large, high-capacity network with a substantial number of parameters, usually well-trained and well-generalized in the target domain. A common challenge in KD is the capacity gap – the discrepancy between the complex teacher and the, student that can hinder effective knowledge transfer. In the context of resource-constrained applications, this gap becomes even more pronounced, as student models must be extremely compact to meet strict memory and computational limits. Unlike conventional KD setups where moderate capacity differences are acceptable, our targeted scenario inherently

TABLE 1. Summary of Knowledge Distillation methods used in this study along with their categories, and the nature of transferred knowledge.

Method Acronym	Method Name	Category	What is Transferred
Vanilla KD [2]	Knowledge Distillation	Response-Based	The teacher's final layer raw outputs (soft targets).
FitNet [39]	Hints for Thin Deep Nets	Feature-Based	Intermediate feature maps from the teacher's hidden layers.
AT [40]	Attention Transfer	Feature-Based	Spatial attention maps, which are summaries of feature maps, to guide where the student "looks".
NST [41]	Neuron Selectivity Transfer	Feature-Based	The statistical properties (via Maximum Mean Discrepancy) of neuron activations in intermediate layers.
VID [42]	Variational Information Distillation	Feature-Based	A variational approximation of the mutual information between teacher and student feature representations.
FT [43]	Factor Transfer	Feature-Based	Paraphrased representations of the teacher's feature maps, learned via factorization, for the student to mimic.
RKD [44]	Relational Knowledge Distillation	Relation-Based	The relationships between different data samples, specifically the distance and angle between their feature vectors in the teacher's embedding space.
SP [45]	Similarity-Preserving KD	Relation-Based	The pairwise similarity matrix of a batch of inputs, ensuring that inputs that are similar in the teacher's space are also similar in the student's space.
CC [46]	Correlation Congruence	Relation-Based	The correlation matrix between instances in a batch, capturing inter-sample relationships at a structural level.
PKT [47]	Probabilistic Knowledge Transfer	Relation-Based	The probability distribution of the feature vectors, thereby transferring structural knowledge of the embedding space by matching these distributions.
CRD [24]	Contrastive Representation Distillation	Relation-Based	Structural knowledge from the teacher's representations using a contrastive loss, which explicitly models the relationships between positive and negative samples.

requires dealing with severe teacher–student disparities. To systematically study this effect, we perform experiments in varying degrees of capacity gaps to assess their impact on student performance under realistic deployment constraints.

3) Heterogeneity in Teacher and Student architectures

Popular lightweight convolutional neural network architectures such as the MobileNet, ShuffleNet, SqueezeNet, EfficientNet, and MCUNet families are specifically designed for optimal performance on resource-constrained hardware platforms. In contrast, traditional architectures like ResNet and similar high-capacity networks are typically employed as teacher models due to their larger parameter counts and strong convergence across a wide range of deep learning tasks. Consequently, the KD setup considered in this work involves heterogeneous teacher–student pairs, where the teacher and student belong to fundamentally different CNN architectural families.

C. DATASETS

1) CIFAR-100

The CIFAR-100 dataset is a widely used benchmark in computer vision for developing and testing machine learning models, particularly for fine-grained object recognition. The dataset consists of 60,000 32x32 pixel color images evenly distributed across 100 distinct classes. The dataset is split into 50,000 images for training and 10,000 for testing, with each class containing exactly 600 images (500 for training and 100 for testing). A key feature of CIFAR-100 is its hierarchical structure. The 100 fine-grained classes are grouped into 20 broader "superclass" categories each containing five related classes.

2) State Farm Distracted Driver Detection dataset

The State Farm Distracted Driver Detection dataset is a cornerstone resource in the field of driver safety and computer vision, created to spur the development of models that can automatically classify a driver's behavior from dashboard camera images. Released as part of a *Kaggle* competition, the dataset consists of total 22,424 labeled 2D color images for training, captured from the perspective of a camera mounted inside a vehicle. Fig. 3 shows sample images, one per class, from the dataset. These images are categorized into ten distinct classes: one representing safe, attentive driving (c0) and nine representing various common distractions. These distractions include phone use (texting or talking with either the left or right hand), operating the radio, drinking, reaching behind, doing hair and makeup, and talking to a passenger. A critical aspect of the dataset's design is that the training and test sets are split by driver, meaning images of a single individual appear in only one of the sets, which challenges models to generalize across different people and vehicle interiors.

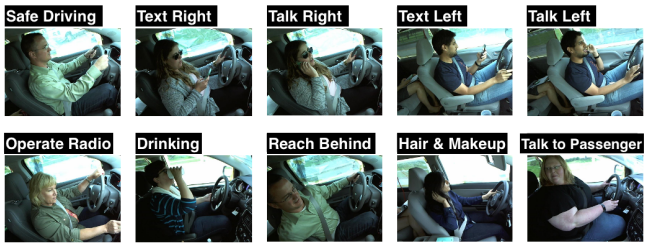

FIGURE 3. Sample images from the State Farm Distracted Driver Detection dataset.

3) CIFAR-100-C

The CIFAR-100-C dataset is a corruption-augmented version of the original CIFAR-100 image classification benchmark. The dataset is created to evaluate the robustness of the models under common image degradations. Specifically, it takes the original 10,000 test images from CIFAR-100 an applies 19 different corruption types that are categorized into noise, blur, weather and digital effects at 5 increasing severity levels, resulting in total of 950,000 corrupted test samples. These corruptions are chosen to simulate real-world degradations and distribution shifts that can degrade the model performance but are often under-explored in standard benchmarking.

4) 100-Driver dataset

The 100-Driver dataset is another dataset built for distracted driver monitoring. This dataset was created considering the limitations of the previous datasets in terms of sample size and environmental variations. The dataset consists of 470K images of diverse postures captured by 4 cameras observing 100 drivers for 79 hours using 5 vehicles. The images show different perspectives from the 4 cameras that were installed in 4 different positions observing the driver. We selected all 27,086 day time images taken by camera number 4 (Cam4) that has similar perspective and highest visual similarity with the State Farm Distracted Driver Detection dataset to maintain an acceptable domain shift for robustness evaluation. The 100-Driver dataset has 18 classes showing different activities of the driver. For consistency we chose the 8 classes in common with the State Farm Distracted Driver Detection dataset, namely safe driving, texting (right), talking on the phone (right), texting (left), talking on the phone (left), operating the radio, reaching behind, and talking to passenger. Fig. 4 shows sample images from the dataset with one example per class.

FIGURE 4. Sample images from the 100-Driver dataset.

D. EVALUATION STRATEGY

In this section, we define the two primary evaluation criteria used in this work. First, we measure overall performance on subsets of the primary datasets – CIFAR-100 and the State Farm Distracted Driver Detection dataset. Second, we evaluate the robustness of student models distilled with different KD techniques (across KD categories) on secondary but related datasets – CIFAR-100-C and 100-Driver.

For all experiments, ResNet32x4 and ResNet110 [51] are used as teacher models, while the student models, ordered from largest to smallest, are MobileNetV2 (α=0.5) [52], ShuffleNetV2 (α=0.2) [53], and MobileNetV2 (α=0.1), where the width multiplier α uniformly scales the number of channels across convolutional layers to trade accuracy for computational cost and parameter count.For the remainder of this work, these student models are termed MobileNetV2 (0.5), ShuffleNetV2 (0.2), and MobileNetV2 (0.1) or MobileNetV2_tiny, respectively. All vanilla teachers and vanilla students are first trained on the primary datasets. The teacher weights are then frozen and student models are distilled using various KD techniques, with each distilled student compared against its corresponding vanilla student using overall F1 scores.

1) Performance on the primary datasets

We first evaluate the performance of vanilla teacher and student models on CIFAR-100 and the State Farm Distracted Driver Detection dataset by reporting overall F1 scores. Specifically, we use the CIFAR-100 validation split (10,000 samples) and a 10,000 sample test subset from the State Farm Distracted Driver Detection dataset. Training hyperparameters and other experimental details are provided in Section IV. We then conduct a comparative analysis over six combinations (two teachers $\times$ three students), recording the F1 scores for each model–dataset pair.

We also compute the total number of *undistillable classes* for every KD technique, using the vanilla student as a reference. A class is counted as undistillable for a given KD method if the distilled student's per-class F1 score does not exceed the vanilla student's per-class F1 on the same evaluation subset. In this work, the motivation behind the calculation of undistillable classes is to provide additional insights beyond overall F1 scores – specifically, to understand how different KD techniques influence the student's ability to learn and transfer discriminative knowledge from the teacher. This analysis helps illustrate how each KD method affects the student's capacity to distinguish between classes, ultimately offering a more detailed perspective on the quality of knowledge transfer.

2) Robustness on unseen data

As discussed in Section I, robustness under distributional and domain shift is essential for real-time, resource-constrained applications. For this evaluation, we take all models trained/distilled on the primary datasets and test them without any fine-tuning on CIFAR-100-C and a subset of the 100-Driver dataset. These datasets are chosen to introduce substantial but relevant perturbations and domain differences,

challenging enough to approximate real-world deployment conditions while remaining close enough to avoid an unfair "out-of-scope" evaluation. This setup allows us to assess whether KD-trained compact students retain accuracy under shifts that are likely to be encountered in practice.

IV. EXPERIMENTAL DETAILS

We use the PyTorch framework [54] to carry out all the experiments in this study with torch version 2.0.1. All experiments were carried out on a host machine with a single NVIDIA GeForce RTX 4090 GPU with CUDA version 12.2. The codebase for all our experiments is built on the github RepDistiller [55]. Following this benchmark setup in [24], we select representative and widely used KD techniques spanning the three standard distillation categories: response/logit-based, feature-based, and relation-based. Rather than exhaustively covering all published variants, our goal is to include strong baselines from each category that are commonly adopted in prior KD literature and compatible with our teacher–student architectures. This ensures a balanced comparison across different forms of transferred knowledge, while keeping the evaluation practical for ultra-compact student models.

All KD methods are trained using the default loss weights and temperature settings provided in the RepDistiller benchmark, following the original implementations. These values are kept fixed across all experiments. While some methods may benefit from task-specific tuning, our goal is to compare different KD categories under a unified training protocol rather than optimize each approach individually.

All teacher and non-distilled student models are first trained on the CIFAR-100 and State Farm Distracted Driver Detection dataset for 240 epochs. We used a batch size of 64 for training on CIFAR-100 and 32 for State Farm Distracted Driver Detection dataset. The model input image resolution for CIFAR-100 is the original 32×32, and for State Farm Distracted Driver Detection dataset, we downsample the original 640×480 to 224×224. During the training we perform augmentation on the images using PyTorch transforms. We perform RandomCrop(size=32, padding=4) and RandomHorizontalFlip on CIFAR-100 and normalize the images using the mean and standard deviation of the overall distribution of the dataset. On the State Farm Distracted Driver Detection dataset, we perform a RandomHorizontalFlip and normalize the images with ImageNet mean and standard deviation [56]. We begin the training with an initial learning rate of 5e-2 for all the teacher models and 1e-2 for all the student models with learning rate decay after epochs 150, 180, 210 with a decay rate of 0.1. The SGD optimizer is used for all training with a weight decay of 5e-4 and a momentum of 0.9. Note that the same hyperparameter and training setting for performing knowledge distillation are used on all student models.

The total number of parameters for all teacher and student models, including both the backbone and the final classifier for the 10-class State Farm Distracted Driver Detection dataset, is summarized in Table 2. The models are ordered from the largest teacher to the smallest student to illustrate their relative capacities. Table 3 further reports the capacity gaps between each teacher–student pair by comparing their total parameter counts, highlighting the capacity gap between compact student models and standard teacher models as discussed in Section III-B2.

TABLE 2. Total parameter counts (in millions) for all teacher and student models used in this study, including backbone and classifier for the 10-class State Farm Distracted Driver Detection dataset. Models are ordered from the largest teacher to the smallest student.

Model	Parameters (M)
ResNet32x4	7.41
ResNet110	1.73
MobileNetV2 (0.5)	0.70
ShuffleNetV2 (0.2)	0.17
MobileNetV2 (0.1)	0.08

TABLE 3. The difference in the total parameters showing capacity gap between the teacher and student models used in this study.

Student	Teacher	Capacity gap (×)
MobileNetV2 (0.5)	ResNet32x4	10.6
MobileNetV2 (0.5)	ResNet110	2.5
ShuffleNetV2 (0.2)	ResNet32x4	42.5
ShuffleNetV2 (0.2)	ResNet110	9.9
MobileNetV2 (0.1)	ResNet32x4	90.4
MobileNetV2 (0.1)	ResNet110	21.1

V. RESULTS AND DISCUSSIONS

A. OVERALL ACCURACY OF DIFFERENT KD METHODS ON CIFAR-100

1) F1 Score

First, we compare the performance of various knowledge distillation (KD) techniques on the CIFAR-100 dataset. Table 4 presents the results obtained using two teacher models and three student models, as described in section IV. The table reports the overall F1 score and the total number of undistillable classes for each method. A detailed examination of the results shows that the CRD and CRD+KD techniques, in the relation-based category, consistently achieve the highest F1 scores and the least number of undistillable classes. In the table, the best results are shown in bold, and the second-best are underlined.

A closer analysis of the teacher-student model pairs reveals a clear trend: the relative improvement in models' accuracy is greatest when the difference in model capacity between teacher and student is small, as observed in the ResNet32x4 and MobileNetV2 (0.5) pairing. This trend is further illustrated in Fig. 5, which visualizes the F1 score improvement over the corresponding vanilla student across all KD categories and student model sizes. In contrast, the smallest performance gain occurs when the capacity gap between teacher and student is largest. This suggests that when the teacher model is substantially larger and more complex than the student, the student struggles to effectively extract and transfer the distilled

knowledge regardless of the specific KD technique or category used.

2) Undistillable classes

We further evaluate the total number of undistillable classes by comparing the F1 scores per-class of each KD technique with those of the vanilla student model. For the CIFAR-100 dataset, we plot the performance changes per-class when distilling from ResNet32x4 to the smallest student (model with least number of total parameters) – MobileNetV2 (0.1) using the best-performing KD technique, CRD. Fig. 6 illustrates this comparison, with class names on the x-axis and the change in the F1 score on the y-axis, ordered from the highest improvement to the highest drop. This visualization provides supplementary evidence that although the overall improvement in the F1 score achieved by CRD over the vanilla student is modest, the student distilled using CRD demonstrates enhanced knowledge transfer through improved class distinction. The total number of undistillable classes for CRD is 35 – the lowest among all evaluated methods. Furthermore, the superclass ''reptiles'' exhibits the highest improvement of 7.24%, suggesting that the performance gains achieved by CRD are systematic rather than random across classes.

B. MODEL ROBUSTNESS USING DIFFERENT KD METHODS ON CIFAR-100-C

The motivation for evaluating performance on CIFAR-100-C is to examine the robustness of the student models following knowledge distillation. CIFAR-100-C introduces a diverse set of severity-controlled corruptions and augmentations, producing input images that simulate distribution shifts. Table 5 summarizes the robustness performance of different KD techniques in combinations of selected students and teachers, where robustness is quantified using the F1-score. The selected student architectures are designed for deployment on edge devices operating in resource-constrained environments. Consequently, they are expected to exhibit enhanced generalization capabilities compared to their vanilla counterparts. Given the low resolution of the CIFAR-100 images, the corruptions and augmentations introduced in CIFAR-100-C create a challenging benchmark with a notable domain shift, particularly under high levels of augmentation severity.

As illustrated in Fig. 7, in line with the trends observed in the accuracy evaluation, relation-based methods, particularly CRD, demonstrate superior performance in terms of robustness. However, this behavior does not extend to the smallest student model, MobileNetV2 (0.1), which exhibits the largest capacity gap relative to teacher networks. Although relation-based distillation effectively brings robustness to the larger student models, it fails to make MobileNetV2 (0.1) robust under the distribution shifts introduced by CIFAR-100-C. This suggests that the excessive capacity gap limits the ultra-compact student from effectively generalizing. Although VID yields the highest F1 score for the smallest student model, the performance differences between VID, CC, CRD, and especially the vanilla model remain marginal. Overall, these results underscore the critical role of the capacity gap in determining the robustness capability of ultra-compact student models.

C. OVERALL ACCURACY OF DIFFERENT KD METHODS ON STATE FARM DISTRACTED DRIVER DETECTION DATASET

1) F1 Score

In this experiment, the evaluation is extended to a more realistic application scenario – driver monitoring system, using the State Farm Distracted Driver Detection dataset. As illustrated in Fig. 3 in Section III-C, this dataset presents substantial complexity due to the subtle inter-class variations, where only a small number of pixels differ between classes while the majority of the image remains unchanged (e.g., talking on the phone versus texting while driving). The dataset comprises of total 22,424 images, and given both its limited size and the fine-grained class distinctions, models – particularly larger teacher networks, tend to become overconfident on the training set. Consequently, the larger teacher model, ResNet32×4, exhibited signs of overfitting, therefore only the smaller teacher network, ResNet110, was used for evaluations.

Table 6 presents the overall F1 scores and the total number of undistillable classes for ResNet110 paired with three student models, across all KD techniques. A closer examination of the results reveals that the standard response-based knowledge distillation method (KD) achieves the best overall performance among teacher–student pairs. Specifically, it achieves the highest F1 scores and only a single undistillable class on the State Farm Distracted Driver Detection test set for MobileNetV2 students. For the ShuffleNetV2 student, the CRD+KD technique produces the highest F1 score, also with the total of only one undistillable class. Although the standard KD method demonstrates the strongest overall results, the relation-based approaches still exhibit consistently higher average performance compared to feature-based methods. In particular, the CRD variants, which achieved superior results on CIFAR-100, also deliver performance improvements over the vanilla student on the State Farm Distracted Driver Detection dataset.

D. MODEL ROBUSTNESS USING DIFFERENT KD METHODS ON 100-DRIVER DATASET

To further assess the robustness of the KD techniques for compact student models, we extend the evaluation by testing models trained on the State Farm Distracted Driver Detection dataset on a completely unseen dataset, the 100-Driver dataset. Fig. 4 shows sample images from this dataset. Although the camera perspectives are visually similar to those in the State Farm Distracted Driver Detection dataset images, a significant domain shift is evident due to factors such as the wide-angle fisheye lens distortion and differences in illumination conditions.

Table 7 reports the F1 scores, which are used to quantify robustness, of both vanilla and distilled models on the unseen

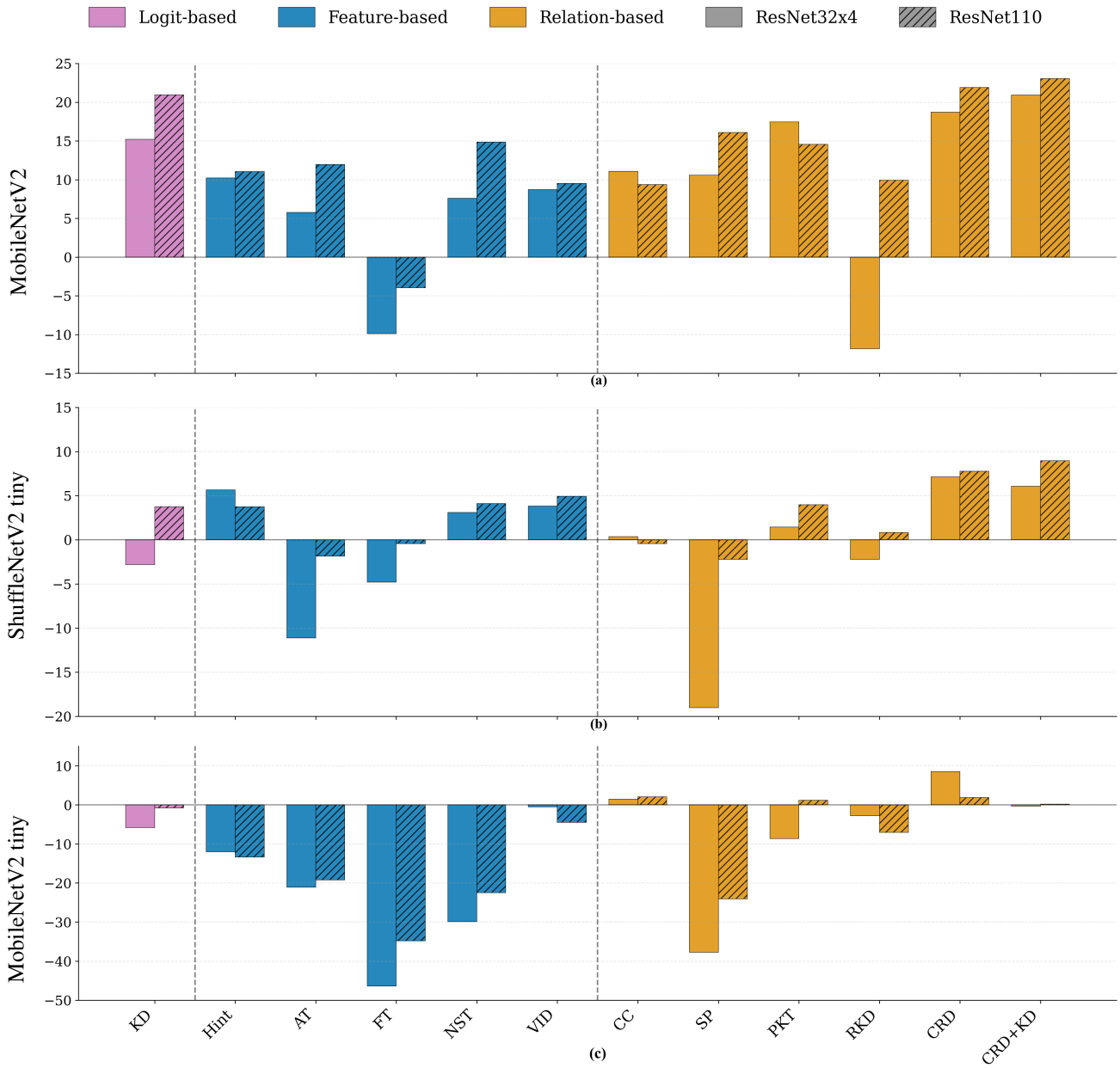

FIGURE 5. Percentage change in macro F1 score relative to the vanilla (undistilled) student on the CIFAR-100 validation set, grouped by KD category: logit-based, feature-based, and relation-based (separated by dashed vertical lines). Solid bars indicate ResNet32×4 as teacher; hatched bars indicate ResNet110. Positive values reflect improvement over the baseline student; negative values reflect degradation. Student models decrease in capacity from top to bottom: (a) MobileNetV2 (0.5) as student, (b) ShuffleNetV2 (0.2) as student, (c) MobileNetV2 (0.1) as student.

100-Driver dataset. Given the inherent challenges in training models on the State Farm Distracted Driver Detection dataset and the increased difficulty introduced by the 100-Driver dataset, the results show that both model variants struggle to generalize effectively. Nevertheless, several noteworthy patterns emerge from the relative comparisons of F1 scores. Among the vanilla students, the smallest model achieves the highest F1 score, suggesting that its lower capacity may allow it to learn the State Farm Distracted Driver Detection dataset features in a more balanced manner, avoiding overconfidence, a behavior observed in larger models, and consequently exhibiting better generalization on the unseen 100-Driver dataset.

An interesting trend can also be observed in relation to model capacity discrepancies. For the largest student model, MobileNetV2 (0.5), where the vanilla performance was the lowest and the teacher–student capacity gap was minimal, knowledge distillation resulted in a substantial performance gain. Specifically, the highest-performing technique, Similarity Preserving (SP), increased the F1 score from **7.23** to

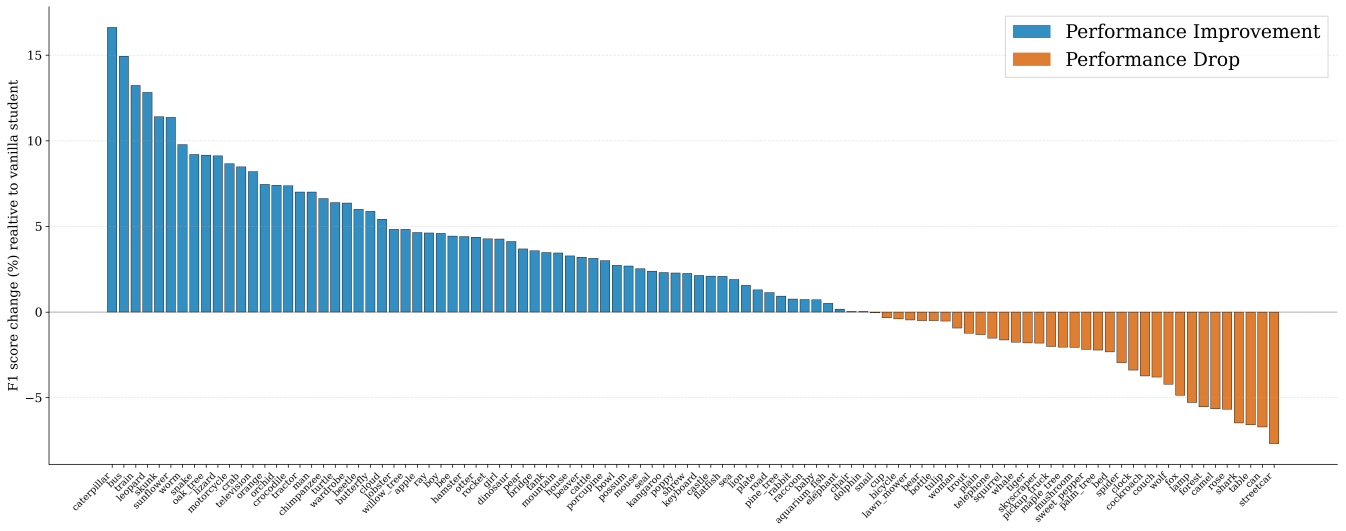

FIGURE 6. **Per-class F1 score change (%) when distilling from ResNet32x4 to MobileNetV2 (0.1) using CRD on the CIFAR-100 validation set. Classes are sorted from highest improvement to highest degradation relative to the standalone student. Blue bars indicate distillable classes (improvement due to distillation), while orange bars indicate undistillable classes where knowledge distillation degrades performance.**

TABLE 4. **Performance comparison of different knowledge distillation techniques across three student models and two teacher models on the CIFAR-100 Val set.**

KD Method	ResNet32x4 (Teacher)			ResNet110 (Teacher)		
	MobileNetV2 (0.5)	ShuffleNetV2 (0.2)	MobileNetV2 (0.1)	MobileNetV2 (0.5)	ShuffleNetV2 (0.2)	MobileNetV2 (0.1)
Vanilla Teacher	68.00	68.00	68.00	61.53	61.53	61.53
Vanilla Student	46.14	50.86	23.72	46.14	50.86	23.72
KD	53.15 | 05	49.43 | 62	22.33 | 58	55.81 | 04	52.76 | 34	23.53 | 55
Hint	50.85 | 19	53.74 | 28	20.86 | 81	51.24 | 13	52.77 | 31	20.54 | 67
AT	48.80 | 27	45.20 | 86	18.72 | 82	51.66 | 13	49.92 | 53	19.16 | 85
FT	41.58 | 79	48.42 | 67	12.72 | 99	44.31 | 68	50.64 | 44	15.46 | 96
NST	49.64 | 23	52.44 | 31	16.62 | 95	52.99 | 12	52.95 | 29	18.38 | 81
VID	50.16 | 16	52.80 | 31	23.59 | 53	50.53 | 14	53.37 | 27	22.65 | 67
CC	51.25 | 10	51.05 | 44	24.07 | 48	50.46 | 11	50.64 | 50	**24.21** | 50
SP	51.02 | 12	41.19 | 97	14.77 | 97	53.56 | 07	49.73 | 51	18.01 | 90
PKT	54.19 | 06	51.60 | 41	21.67 | 69	52.86 | 02	52.88 | 27	24.00 | 51
RKD	40.68 | 91	49.73 | 60	23.06 | 50	50.72 | 12	51.28 | 45	22.04 | 64
CRD	54.77 | 01	**54.48** | 16	**25.74** | 35	56.25 | 03	54.81 | 10	24.17 | 48
CRD+KD	**55.79** | 04	53.95 | 17	23.63 | 50	**56.77** | 02	**55.42** | 10	23.76 | 48

All values are reported as **F1 Score** (macro). The values are shown as "score | total undistillable classes". The maximum F1 scores are shown in bold and the second highest are underlined. Horizontal separator lines are used to distinguish KD categories, ordered from top to bottom as response-based, feature-based, and relation-based methods.

18.98 which is approximately 2.6 ×. In contrast, for the smallest student model, where the vanilla student already achieved the highest performance and the teacher–student capacity gap was the largest, the improvement achieved by the best-performing technique, Attention Transfer (AT), was comparatively modest of approximately 1.4 ×. Overall, when averaged across all methods within each category of KD, relation-based techniques achieved the highest average F1 scores, with CRD variants – particularly for MobileNet-family student models, consistently achieving the second-best performance with only marginal differences compared to the top-performing methods.

VI. CONCLUSIONS AND FUTURE WORK

In resource-constrained TinyML applications, maintaining both model accuracy and low latency is a fundamental challenge. Models intended for deployment on low-spec platforms must not only be compact but also robust to variations in data distributions since real-world input often differs from the training data. Hence, Knowledge Distillation (KD) is a good choice for such scenarios. For real-time applications such as Driver Monitoring Systems, where latency is critical, extremely lightweight models are required. Consequently, the KD setup for such scenarios naturally involves high-capacity teacher networks and compact student models. Accordingly, our evaluation centers on understanding how knowledge distillation behaves when student capacity is severely limited.

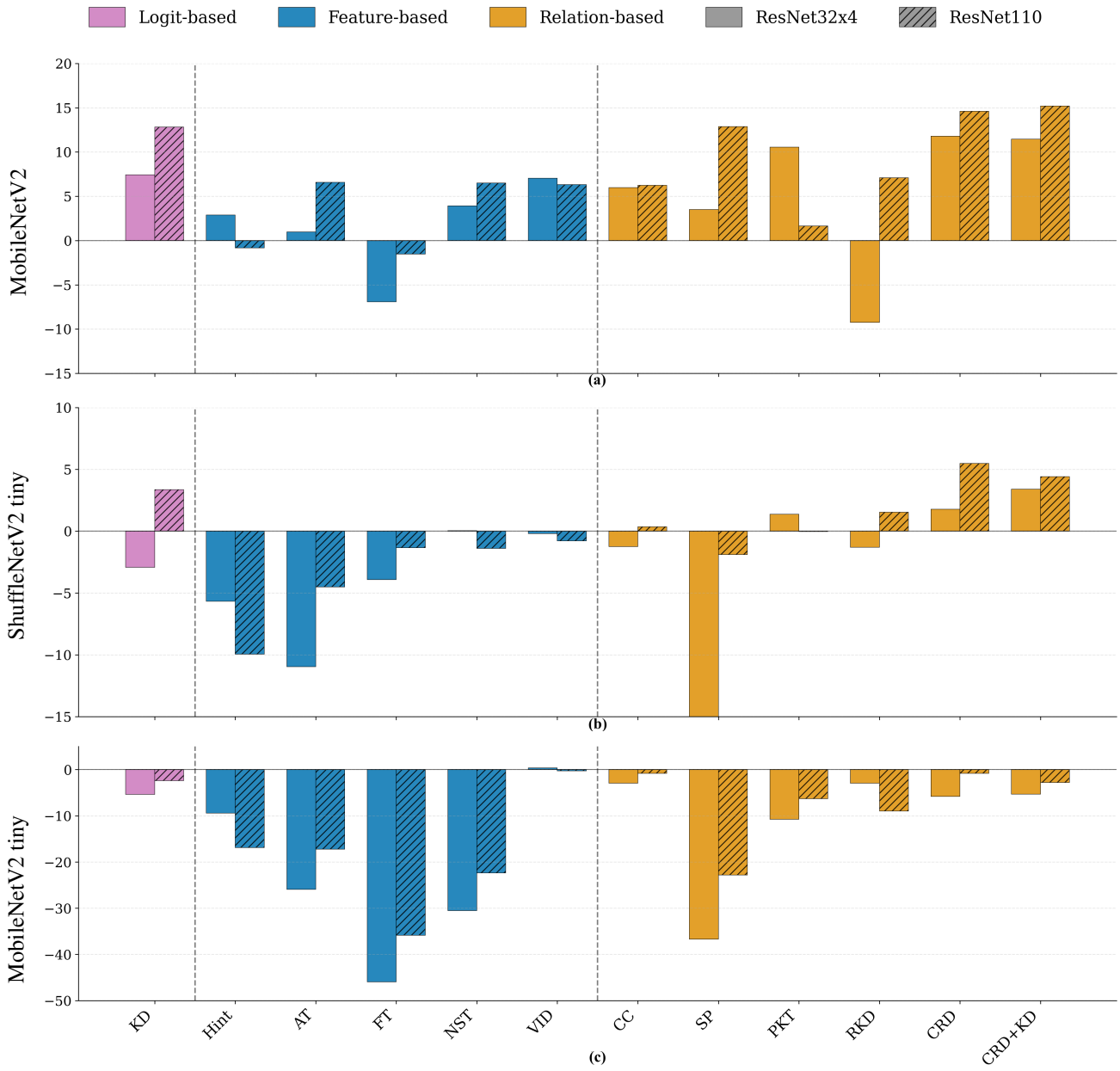

FIGURE 7. Percentage change in macro F1 score relative to the vanilla (undistilled) student on the CIFAR-100-C corruption benchmark, grouped by KD category: logit-based, feature-based, and relation-based (separated by dashed vertical lines). Solid bars indicate ResNet32×4 as teacher; hatched bars indicate ResNet110. Positive values reflect robustness improvement over the baseline student; negative values reflect degradation. Student models decrease in capacity from top to bottom: (a) MobileNetV2 (0.5) as student, (b) ShuffleNetV2 (0.2) as student, (c) MobileNetV2 (0.1) as student.

We assume that the selected ultra-compact models meet basic MCU memory requirements, and focus on comparing KD methods in terms of accuracy and robustness, rather than reporting hardware-level metrics.

From our experiments involving ResNet-family teachers and MobileNet and ShuffleNet-family students on the CIFAR-100 benchmark dataset, we observed that relation-based KD methods, particularly Contrastive Representation Distillation (CRD), achieved superior overall accuracy. These methods also performed strongly on the CIFAR-100-C dataset, which consists of heavily augmented and corrupted samples, indicating their robustness and better generalization to distribution shifts. Additionally, the results highlighted the impact of the teacher–student capacity gap: improvements in F1 score over the vanilla student were most significant when this gap was relatively small, suggesting the importance of identifying an appropriate level of teacher complexity for effective distillation.

In the more challenging and limited State Farm Distracted Driver Detection dataset, the simple response-based vanilla

TABLE 5. Robustness evaluation of knowledge distillation techniques across three student models and two teacher models on the CIFAR-100-C dataset.

KD Method	ResNet32x4 (Teacher)			ResNet110 (Teacher)		
	MobileNetV2 (0.5)	ShuffleNetV2 (0.2)	MobileNetV2 (0.1)	MobileNetV2 (0.5)	ShuffleNetV2 (0.2)	MobileNetV2 (0.1)
Vanilla Teacher	50.22	50.22	50.22	47.16	47.16	47.16
Vanilla Student	41.60	42.30	**26.61**	41.60	42.30	**26.61**
KD	44.69	41.06	25.18	46.93	43.72	25.97
Hint	42.80	39.90	24.10	41.25	38.09	22.12
AT	42.01	37.66	19.71	44.34	40.39	22.03
FT	38.72	40.64	14.38	40.96	41.73	17.07
NST	43.23	42.32	18.49	44.30	41.71	20.65
VID	44.53	42.22	**26.71**	44.23	41.97	**26.53**
CC	44.09	41.77	25.83	44.20	42.45	26.39
SP	43.06	35.05	16.84	46.96	41.5	20.54
PKT	45.99	42.88	23.75	42.29	42.29	24.93
RKD	37.75	41.75	25.82	44.55	42.95	24.22
CRD	**46.51**	**43.05**	25.06	47.68	**44.62**	26.39
CRD+KD	46.37	43.74	25.19	**47.92**	44.17	25.87

All values are reported as **F1 Score** (macro). The values are shown as "score | total undistillable classes". The maximum F1 scores are shown in bold and the second highest are underlined. Horizontal separator lines are used to distinguish KD categories, ordered from top to bottom as response-based, feature-based, and relation-based methods.

TABLE 6. Performance comparison of different knowledge distillation techniques across three student models and ResNet110 teacher model on the State Farm Distracted Driver Detection dataset.

KD Method	ResNet110 (Teacher)		
	MobileNetV2 (0.5)	ShuffleNetV2 (0.2)	MobileNetV2 (0.1)
Vanilla Teacher	84.88	84.88	84.88
Vanilla Student	81.43	76.87	40.23
KD	**87.59** | 1	85.52 | 1	**70.00** | 1
Hint	68.30 | 9	51.63 | 10	25.88 | 9
AT	77.27 | 7	75.76 | 5	46.27 | 6
FT	70.89 | 8	78.28 | 3	33.93 | 7
NST	82.41 | 4	79.18 | 4	42.45 | 6
VID	79.32 | 6	82.63 | 2	35.03 | 6
CC	76.54 | 7	54.27 | 9	34.56 | 8
SP	85.78 | 2	84.32 | 1	44.05 | 5
PKT	83.51 | 2	83.91 | 2	42.52 | 3
RKD	50.78 | 9	61.69 | 6	27.07 | 8
CRD	72.80 | 5	77.83 | 4	53.85 | 3
CRD+KD	83.18 | 3	**87.77** | 1	44.69 | 5

All values are reported as **F1 Score** (macro). The values are shown as "score | total undistillable classes". The maximum F1 scores are shown in bold and the second highest are underlined. Horizontal separator lines are used to distinguish KD categories, ordered from top to bottom as response-based, feature-based, and relation-based methods.

KD method showed the highest accuracy among all, especially for the smallest student model with the largest capacity gap. Relation-based methods such as CRD+KD and SP also demonstrated competitive performance. Taking the challenge even further, we evaluated the models trained on the State Farm Distracted Driver Detection dataset on a different dataset for the same application, 100-Driver, and observed that CRD+KD and SP again achieved the highest robustness, surpassing the vanilla student models.

These findings provide valuable insights into the most effective KD category for real-time, resource-constrained applications with highly compact CNNs, and clarify how varying capacity gaps influence distillation outcomes. However, the experiments are limited to heterogeneous CNN architectures and do not explore homogeneous or cross-architecture configurations such as CNN-to-ViT distillation. Future work can extend this direction to better understand architectural compatibility and evaluate KD effectiveness across diverse model families. Moreover, the observed improvement in robustness on the 100-Driver dataset suggests that KD may also act as a regularizer, enhancing the generalization of compact models under challenging, unseen real-world conditions, which can be explored in future studies.

ACKNOWLEDGEMENT

We would like to express our gratitude to Luna Systems for their invaluable support throughout the course of this research.

TABLE 7. Robustness evaluation of knowledge distillation techniques across three student models and ResNet110 teacher model on the 100-Driver dataset.

KD Method	ResNet110 (Teacher)		
	MobileNetV2 (0.5)	ShuffleNetV2 (0.2)	MobileNetV2 (0.1)
Vanilla Teacher	17.56	17.56	17.56
Vanilla Student	7.23	7.65	9.74
KD	14.93	10.12	12.26
Hint	10.90	12.28	12.34
AT	8.08	8.52	**13.61**
FT	8.26	12.56	7.86
NST	15.73	11.41	8.86
VID	11.40	<u>14.32</u>	6.76
CC	11.40	9.86	9.32
SP	**18.98**	**14.72**	10.60
PKT	14.90	11.93	11.65
RKD	8.28	6.83	10.19
CRD	<u>17.66</u>	10.05	11.52
CRD+KD	9.45	11.10	<u>13.07</u>

All values are reported as **F1 Score** (macro). The values are shown as "score | total undistillable classes". The maximum F1 scores are shown in bold and the second highest are underlined. Horizontal separator lines are used to distinguish KD categories, ordered from top to bottom as response-based, feature-based, and relation-based methods.

REFERENCES

[1] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *International journal of computer vision*, vol. 129, no. 6, pp. 1789–1819, 2021.

[2] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[3] B. Huang, M. Chen, Y. Wang, J. Lu, M. Cheng, and W. Wang, "Boosting accuracy and robustness of student models via adaptive adversarial distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 24 668–24 677.

[4] Z. Xue, Z. Gao, S. Ren, and H. Zhao, "The modality focusing hypothesis: Towards understanding crossmodal knowledge distillation," *arXiv preprint arXiv:2206.06487*, 2022.

[5] F. Huo, W. Xu, J. Guo, H. Wang, and S. Guo, "C2kd: Bridging the modality gap for cross-modal knowledge distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 16 006–16 015.

[6] H. Wang, C. Ma, J. Zhang, Y. Zhang, J. Avery, L. Hull, and G. Carneiro, "Learnable cross-modal knowledge distillation for multi-modal learning with missing modality," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2023, pp. 216–226.

[7] S. Du, S. You, X. Li, J. Wu, F. Wang, C. Qian, and C. Zhang, "Agree to disagree: Adaptive ensemble knowledge distillation in gradient space," *advances in neural information processing systems*, vol. 33, pp. 12 345–12 355, 2020.

[8] W. Son, J. Na, J. Choi, and W. Hwang, "Densely guided knowledge distillation using multiple teacher assistants," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9395–9404.

[9] L. Dutta and S. Bharali, "Tinyml meets iot: A comprehensive survey," *Internet of Things*, vol. 16, p. 100461, 2021.

[10] C. Kaundanya, P. Cesar, B. Cronin, A. Fleury, M. Liu, and S. Little, "Using attention mechanisms in compact cnn models for improved micromobility safety through lane recognition," in *Proceedings of the 10th International Conference on Vehicle Technology and Intelligent Transport Systems*. SCITEPRESS - Science and Technology Publications, 2024.

[11] ——, "Enhancing small object detection in resource-constrained aras using image cropping and slicing techniques," in *Proceedings of the 20th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 3: VISAPP*, INSTICC. SciTePress, 2025, pp. 570–583.

[12] I. Froiz-Miguez, P. Fraga-Lamas, and T. M. Fernandez-Carames, "Design, implementation, and practical evaluation of a voice recognition based iot home automation system for low-resource languages and resource-constrained edge iot devices: A system for galician and mobile opportunistic scenarios," *IEEE Access*, vol. 11, pp. 63 623–63 649, 2023.

[13] J. Roostaei, Y. Z. Wager, W. Shi, T. Dittrich, C. Miller, and K. Gopalakrishnan, "Iot-based edge computing (iotec) for improved environmental monitoring," *Sustainable computing: informatics and systems*, vol. 38, p. 100870, 2023.

[14] V. Rajapakse, I. Karunanayake, and N. Ahmed, "Intelligence at the extreme edge: A survey on reformable tinyml," *ACM Computing Surveys*, vol. 55, no. 13s, pp. 1–30, 2023.

[15] J. Guo, M. Chen, Y. Hu, C. Zhu, X. He, and D. Cai, "Reducing the teacher-student gap via spherical knowledge disitllation," *arXiv preprint arXiv:2010.07485*, 2020.

[16] Y. Zhu, N. Liu, Z. Xu, X. Liu, W. Meng, L. Wang, Z. Ou, and J. Tang, "Teach less, learn more: On the undistillable classes in knowledge distillation," *Advances in Neural Information Processing Systems*, vol. 35, pp. 32 011–32 024, 2022.

[17] S. I. Mirzadeh, M. Farajtabar, A. Li, N. Levine, A. Matsukawa, and H. Ghasemzadeh, "Improved knowledge distillation via teacher assistant," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 04, 2020, pp. 5191–5198.

[18] C. Wang and Z. Tang, "The staged knowledge distillation in video classification: Harmonizing student progress by a complementary weakly supervised framework," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 8, pp. 6646–6660, 2023.

[19] D. P. Ganta, H. D. Gupta, and V. S. Sheng, "Knowledge distillation via weighted ensemble of teaching assistants," in *2021 IEEE International Conference on Big Knowledge (ICBK)*. IEEE, 2021, pp. 30–37.

[20] A. Krizhevsky, V. Nair, and G. Hinton, "Cifar-10 and cifar-100 datasets," *URl: https://www. cs. toronto. edu/kriz/cifar. html*, vol. 6, no. 1, p. 1, 2009.

[21] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," *arXiv preprint arXiv:1903.12261*, 2019.

[22] State Farm, "State Farm Distracted Driving Dataset," 2016, [Online]. Available: https://www.kaggle.com/c/state-farm-distracted-driver-detection/data. Accessed: Oct. 15, 2025.

[23] J. Wang, W. Li, F. Li, J. Zhang, Z. Wu, Z. Zhong, and N. Sebe, "100-driver: A large-scale, diverse dataset for distracted driver classification," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 7, pp. 7061–7072, 2023.

[24] Y. Tian, D. Krishnan, and P. Isola, "Contrastive representation distillation," *arXiv preprint arXiv:1910.10699*, 2019.

[25] C. Qian, T. Le, and M. Harandi, "A good teacher adapts their knowledge for distillation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 1239–1248.

[26] T. Li, L. Liu, Y. Hu, H. Chen, and S. Chen, "Dual-forward path teacher knowledge distillation: Bridging the capacity gap between teacher and student," *arXiv preprint arXiv:2506.18244*, 2025.

[27] D. Y. Park, M.-H. Cha, D. Kim, B. Han *et al.*, "Learning student-friendly

teacher networks for knowledge distillation," *Advances in neural information processing systems*, vol. 34, pp. 13 292–13 303, 2021.

[28] Y. Liu, Q. Yi, and J. Zeng, "Reducing capacity gap in knowledge distillation with review mechanism for crowd counting," *arXiv preprint arXiv:2206.05475*, 2022.

[29] Y. Xiong, W. Zhai, X. Xu, J. Wang, Z. Zhu, C. Ji, and J. Cao, "Ability-aware knowledge distillation for resource-constrained embedded devices," *Journal of Systems Architecture*, vol. 141, p. 102912, 2023.

[30] M. A. Khan, H. Menouar, R. Hamila, and A. Abu-Dayya, "Crowd counting at the edge using weighted knowledge distillation," *Scientific Reports*, vol. 15, no. 1, pp. 1–16, 2025.

[31] M. Ohamouddou, S. Ohamouddou, A. El Afia, and R. Lasri, "Atms-kd: Adaptive temperature and mixed sample knowledge distillation for a lightweight residual cnn in agricultural embedded systems," *Smart Agricultural Technology*, p. 101617, 2025.

[32] Y. Zhao, S. Li, and X. Feng, "Lightweight remote sensing scene classification on edge devices via knowledge distillation and early-exit," in *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025, pp. 11 862–11 870.

[33] D. Liu, T. Yamasaki, Y. Wang, K. Mase, and J. Kato, "Toward extremely lightweight distracted driver recognition with distillation-based neural architecture search and knowledge transfer," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 1, pp. 764–777, 2022.

[34] W. Li, J. Wang, T. Ren, F. Li, J. Zhang, and Z. Wu, "Learning accurate, speedy, lightweight cnns via instance-specific multi-teacher knowledge distillation for distracted driver posture identification," *IEEE transactions on intelligent transportation systems*, vol. 23, no. 10, pp. 17 922–17 935, 2022.

[35] C. Tanama, K. Peng, Z. Marinov, R. Stiefelhagen, and A. Roitberg, "Quantized distillation: Optimizing driver activity recognition models for resource-constrained environments," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 5479–5486.

[36] H. Ahuja, S. Saurav, S. Srivastava, and C. Shekhar, "Driver drowsiness detection using knowledge distillation technique for real time scenarios," in *2020 IEEE 17th India Council International Conference (INDICON)*. IEEE, 2020, pp. 1–5.

[37] A. Gupta, V. Maan, K. S. Sangwan *et al.*, "Optimized human drowsiness detection using spatial and temporal features of eye and mouth," in *2023 10th International Conference on Signal Processing and Integrated Networks (SPIN)*. IEEE, 2023, pp. 866–871.

[38] C. Yang, X. Yu, Z. An, and Y. Xu, "Categories of response-based, feature-based, and relation-based knowledge distillation," in *Advancements in knowledge distillation: towards new horizons of intelligent systems*. Springer, 2023, pp. 1–32.

[39] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "Fitnets: Hints for thin deep nets," 2015. [Online]. Available: https://arxiv.org/abs/1412.6550

[40] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," *arXiv preprint arXiv:1612.03928*, 2016.

[41] Z. Huang and N. Wang, "Like what you like: Knowledge distill via neuron selectivity transfer," *arXiv preprint arXiv:1707.01219*, 2017.

[42] S. Ahn, S. X. Hu, A. Damianou, N. D. Lawrence, and Z. Dai, "Variational information distillation for knowledge transfer," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 9163–9171.

[43] J. Kim, S. Park, and N. Kwak, "Paraphrasing complex network: Network compression via factor transfer," *Advances in neural information processing systems*, vol. 31, 2018.

[44] W. Park, D. Kim, Y. Lu, and M. Cho, "Relational knowledge distillation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3967–3976.

[45] F. Tung and G. Mori, "Similarity-preserving knowledge distillation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1365–1374.

[46] B. Peng, X. Jin, J. Liu, D. Li, Y. Wu, Y. Liu, S. Zhou, and Z. Zhang, "Correlation congruence for knowledge distillation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 5007–5016.

[47] N. Passalis, M. Tzelepi, and A. Tefas, "Probabilistic knowledge transfer for lightweight deep representation learning," *IEEE Transactions on Neural Networks and learning systems*, vol. 32, no. 5, pp. 2030–2039, 2020.

[48] J. Lin, W.-M. Chen, Y. Lin, C. Gan, S. Han *et al.*, "Mcunet: Tiny deep learning on iot devices," *Advances in neural information processing systems*, vol. 33, pp. 11 711–11 722, 2020.

[49] A. Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[50] K. Wu, J. Zhang, H. Peng, M. Liu, B. Xiao, J. Fu, and L. Yuan, "Tinyvit: Fast pretraining distillation for small vision transformers," in *European conference on computer vision*. Springer, 2022, pp. 68–85.

[51] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[52] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[53] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "Shufflenet v2: Practical guidelines for efficient cnn architecture design," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 116–131.

[54] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[55] Y. Li, "Repdistiller: A library for knowledge distillation," https://github.com/HobbitLong/RepDistiller, 2019, accessed: Feb. 12, 2025.

[56] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

CHINMAYA KAUNDANYA is a Ph.D. researcher at the Insight Research Ireland Centre for Data Analytics, supervised by Prof. Suzanne Little at Dublin City University. His PhD is funded by the Insight Centre and an AI/Micromobility Company, Luna Systems, based in Dublin, Ireland. His research focuses on the design and optimization of Computer Vision Deep Neural Network models to achieve an effective trade-off between accuracy and latency on resource-constrained edge platforms. His work investigates deployment-oriented model optimization strategies and system-level challenges associated with running vision models on low-spec hardware, particularly for intelligent safety systems in micromobility and automotive applications.

PAULO CÉSAR PEREIRA JÚNIOR received his B.Sc. degree in Computer Science from the Universidade Federal de Santa Catarina (UFSC), Brazil, in 2016, and his M.Sc. degree in Computer Science from UFSC's Graduate Program in Computer Science (PPGCC) in early 2020, with a focus on computer vision. He is currently working as a Computer Vision Engineer at Luna Systems, where he develops machine learning and vision-based solutions for real-world applications. Originally from Brazil and currently based in Ireland, his professional and research interests include computer vision, machine learning, and perception systems, with recent work strongly focused on Advanced Driver Assistance Systems (ADAS) and intelligent sensing technologies.

BARRY CRONIN is CTO at Transpoco, a GPS fleet management and telematics company, a role he's held since 2004. He leads an engineering team through complex technical initiatives including AI platform development, backend modernisation, and security architecture implementation. From 2020-2024, he also served as Technical Manager at Luna Systems, bringing cross-industry technical leadership experience. With deep expertise in fleet management systems, telematics hardware, and software architecture, he balances technical excellence with practical business outcomes. Barry focuses on scaling engineering capabilities, optimising driver scoring systems, and developing strategic roadmaps for OEM integrations and EV capabilities.

ANDREW FLEURY is the CEO & Co-Founder at Luna Systems. Andrew has spent the last two decades exclusively dedicated to the fields of telematics, fleet management and smart vehicle technology. In 2020, he co-founded Luna with the mission of applying computer vision to solve the common safety challenges facing the wide scale adoption of micromobility. The company has built a product portfolio, offering Advanced Rider Assistance Systems (ARAS) to the cycling and motorcycle markets in a bid to support the global shift to sustainable mobility.

MINGMING LIU (SENIOR MEMBER, IEEE) is an Assistant Professor in the School of Electronic Engineering at Dublin City University (DCU). He is also affiliated with the Insight Research Ireland Centre for Data Analytics as a Funded Investigator. He received a B. Eng (1st Hons) in Electronic Engineering from the National University of Ireland Maynooth in 2011 and his PhD in Control Engineering and Decision Science from the Hamilton Institute at the same university in 2015. Prior to DCU, he worked with University College Dublin and IBM Ireland Lab as an applied researcher and EU H2020 project lead. He has published over 70 papers to date including ''IEEE Transactions on Smart Grid'', ''IEEE Transactions on Intelligent Transportation Systems'', ''IEEE Transactions on Automation Science and Engineering'', ''IEEE Transactions on Cloud Computing'', ''IEEE Transactions on Transportation Electrification'', ''IEEE Transactions on Artificial Intelligence'',''IEEE Systems Journal'', ''Sustainable Cities and Society'', ''Pattern Recognition'', ''Applied Energy'', and ''Neural Networks''. He acts as an academic editor for PLOS ONE and an associate editor for the Journal of Information Systems and Operational Research (INFOR). His research interests include control, optimisation and machine learning with applications to IoT, cloud computing, electric vehicles, smart grids, smart transportation, and smart cities.

SUZANNE LITTLE received a Ph.D. from the University of Queensland, Australia, in 2008. She is a Professor in the School of Computing at Dublin City University, Principal Investigator at the Research Ireland Insight Centre for Data Analytics and Co-Director of the RI Centre for Research Training in Artificial Intelligence. Her research interests are in content-based analysis and indexing of digital multimedia, particularly images and video, using computer vision, machine learning, and semantic technologies. Her work spans object and event detection, segmentation, classification, and tracking, with applications in areas including responsible AI, smart cities, transportation, healthcare, and scientific data management.