Punching Above Precision: Small Quantized Model Distillation with Learnable Regularizer

Abdur Rehman abdur@opt-ai.kr
S M A Sharif sharif@opt-ai.kr
Md Abdur Rahaman rahaman@opt-ai.kr
Mohamed Jismy Aashik Rasool rasool@opt-ai.kr
Seongwan Kim swan.kim@opt-ai.kr
Jaeho Lee jaeho.lee@opt-ai.kr
Opt-AI, Seoul, South Korea

Editors: Hung-yi Lee and Tongliang Liu

Abstract

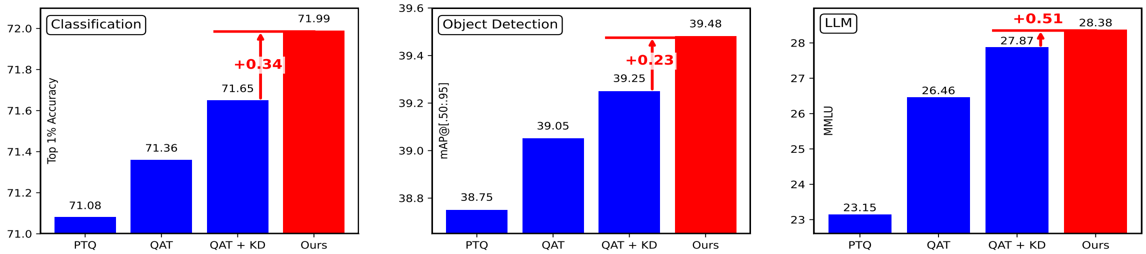
Quantization-aware training (QAT) combined with knowledge distillation (KD) is a promising strategy for compressing Artificial Intelligence (AI) models for deployment on resource-constrained hardware. However, existing QAT-KD methods often struggle to balance task-specific (TS) and distillation losses due to heterogeneous gradient magnitudes, especially under low-bit quantization. We propose Game of Regularizer (GoR), a novel learnable regularization method that adaptively balances TS and KD objectives using only two trainable parameters for dynamic loss weighting. GoR reduces conflict between supervision signals, improves convergence, and boosts the performance of small quantized models (SQMs). Experiments on image classification, object detection (OD), and large language model (LLM) compression show that GoR consistently outperforms state-of-the-art QAT-KD methods. On low-power edge devices, it delivers faster inference while maintaining full-precision accuracy. We also introduce QAT-EKD-GoR, an ensemble distillation framework that uses multiple heterogeneous teacher models. Under optimal conditions, the proposed EKD-GoR can outperform full-precision models, providing a robust solution for real-world deployment.

Keywords: Quantization-aware training; Knowledge distillation; Model compression; Learnable regularizer; Small quantized models

1. Introduction

AI-driven tasks on low-power edge devices are reshaping the next-generation tech ecosystem by enabling critical applications in diverse domains such as healthcare, autonomous systems, and the Internet of Things (IoT) (Rasmussen et al., 2021). However, limited computational and memory resources pose significant challenges for deploying high-performance models in such settings (A Sharif et al., 2024). As a result, compact and optimized models are essential for effective deployment in resource-constrained environments. To address this, various model compression techniques such as quantization, pruning, and KD have been explored; yet, achieving an optimal trade-off between efficiency and performance remains an open challenge.

Among popular strategies for model optimization, QAT has emerged as a powerful technique that enables high-precision floating-point parameters to be converted into low-bit



(a) (b) (c)

Figure 1: QAT-KD with GoR outperforms traditional PTQ, QAT, and QAT+KD methods. (a) classification (top-1 accuracy), (b) object detection (mAP), and (c) large language model (MMLU)

integer representations, significantly reducing model complexity and improving inference efficiency (A Sharif et al., 2024). However, aggressive quantization often results in substantial accuracy degradation due to unavoidable quantization errors and information loss, limiting its practical effectiveness (Pham et al., 2023a). To address this issue, recent approaches have combined QAT with KD, employing a larger teacher model to transfer robust knowledge representations to a smaller quantized student model, thus mitigating quantization-induced accuracy losses (Pham et al., 2023a).

Despite recent advances, QAT-KD methods (Zhu et al., 2023; Zhao and Zhao, 2024) still struggle to balance task-specific (e.g., cross-entropy) and distillation (e.g., KL divergence) losses due to heterogeneous gradient magnitudes and differing optimization dynamics, especially under low-bit quantization. Many recent methods counter these phenomena, relying on extensive hyperparameter tuning to achieve effective convergence (Boo et al., 2021b; Zhuang et al., 2018; Wang et al., 2021; Pham et al., 2023a). Although static or heuristic weighting strategies are commonly used to combine multiple objectives, their lack of adaptability frequently leads to suboptimal learning in SQMs. In contrast, (Zhao and Zhao, 2024) (SQAKD) argue that the TS and KD losses are inherently incompatible. Thus, they propose eliminating TS loss when performing QAT-KD and relying solely on KD guidance instead. Arguably, removing the supervision of TS labels may hinder generalization and degrade performance in real-world scenarios.

The contradictions among existing QAT-KD approaches motivated us to delve deeper into the interaction between TS and KD during QAT. We observe that the conflict between TS and KD is not inherent but rather arises from the absence of an adaptive mechanism to balance them effectively. For instance, completely discarding TS guidance ignores the critical label-driven supervision necessary for semantic alignment and generalization. To address these limitations, we propose GoR, a novel learnable regularization strategy that jointly optimizes both the distillation temperature and loss weighting during training. By adaptively balancing the contributions of TS and KD signals, GoR mitigates quantization-induced performance degradation and fosters better synergy between learning objectives. As shown in Fig. 1, our method consistently outperforms state-of-the-art QAT and QAT-KD baselines, achieving improved generalization and robustness across diverse tasks on low-bit precision (i.e., 8-bit and 4-bit).

We validated across a variety of compact deep architectures and benchmark datasets. Our results demonstrate its strong generalization ability across diverse AI tasks, such as image classification, OD, and LLM compression. Notably, the proposed QAT-KD-GoR significantly enhances real-world performance on edge hardware, offering up to 242% faster inference, making it an ideal solution for deployment in resource-constrained environments. Additionally, we explore the potential of Ensemble Knowledge Distillation (EKD), a strategy that combines knowledge from multiple heterogeneous teacher models. Our findings suggest that combining EKD with the GoR mechanism can further boost the performance of SQM in the quantized regime, even outperforming their full-precision counterparts under optimal conditions. In summary, our key contributions are as follows:

- We introduce GoR, a novel, learnable regularization technique that dynamically optimizes the balance between TS-loss and KD-loss, significantly improving accuracy and convergence for SQM.
- We demonstrate the generalizability of the proposed GoR across diverse AI tasks, including image classification, OD, and LLM. To validate its practical applicability, we evaluate GoR under real-world hardware constraints. Our results show that GoR can enhance the performance of existing methods while introducing negligible additional training parameters.
- We introduce EKD as a practical solution to address the challenge of lacking homogeneous large-scale teacher models by synthesizing knowledge from heterogeneous large-scale models.

2. Related Work

2.1. Quantization-Aware Training

Quantization methods mainly include Post-Training Quantization (PTQ) (Li et al., 2019), which quantizes pre-trained models without retraining but often suffers accuracy loss, and QAT (Krishnamoorthi, 1806), which incorporates quantization during retraining. Early QAT works such as BNN (Courbariaux et al., 2016), XNOR-Net (Rastegari et al., 2016), and DoReFa-Net (Zhou et al., 2016) introduced various scaling techniques, while recent methods employ trainable quantizer parameters to optimize clipping ranges, APoT (Li et al., 2019), DSQ (Gong et al., 2019), and quantization intervals QIL (Jung et al., 2019). Despite improvements, QAT methods vary in effectiveness and lack a unified theory, with MQBench (Li et al., 2021b) showing no single method excels universally, especially at low-bit precision.

2.2. Knowledge Distillation

KD methods, introduced by Hinton et al. (Hinton et al., 2015), minimize the KL divergence between teacher and student softmax outputs, alongside cross-entropy loss, and are typically divided into logit-based and feature-based approaches. Logit-based KD, aligning the output predictions of teacher and student networks, is effective for classification and prediction tasks. In contrast, feature-based KD transfers intermediate representations and aligns well with tasks that require richer spatial information, such as detection (Lightly.ai, 2025). Methods like Channel-Wise Distillation (CWD) (Shu et al., 2021) and Masked Generative Distillation (MGD) (Yang et al., 2022) enhance feature alignment between teacher and student for effective KD on OD. Further research has explored transferring intermediate representations like FSP matrices (Yim et al., 2017), attention maps (Zagoruyko and Komodakis, 2016), internal features (Zhao et al., 2023), modified distillation loss (Zhou et al., 2021). A few methods also utilize multiple teachers for effective distillation. For instance, adaptive multi-teacher (AMT) (Liu et al., 2020) adaptively weights teacher outputs to generate soft targets, and confidence-aware multi-teacher knowledge distillation (CAMKD) (Zhang et al., 2022) estimates per-sample teacher reliability using ground truth to guide label fusion.

2.3. Quantization with Knowledge Distillation

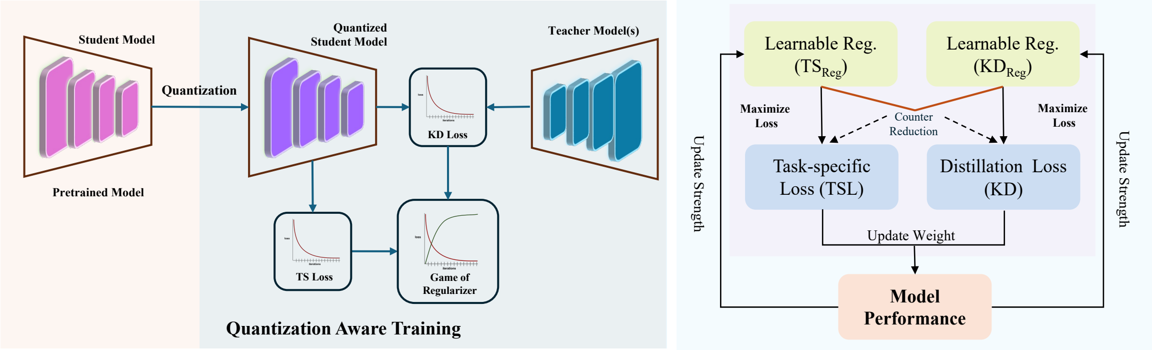
Recent research uses KD to mitigate accuracy drops in low-precision models. (Mishra et al., 2018) introduced Apprentice for ternary-precision and 4-bit networks. QKD (Kim et al., 2019) divides the process into phases for KD and quantization coordination. SPEQ (Boo et al., 2021a) constructs the teacher from the student's parameters with stochastic bit precision. PTG (Zhuang et al., 2018) uses joint training, incremental bit-width reduction, and quantized weight optimization. In QFD (Zhu et al., 2023), a quantized teacher model is used for performing distillation on low-bitwidth networks. Collaborative multi-teacher knowledge distillation (CMT) (Pham et al., 2023b) encourages mutual learning with multiple quantized teachers. SQAKD (Zhao and Zhao, 2024) is a self-supervised framework that argues that using only KD loss is sufficient for KD training, achieving state-of-the-art results. Notably, existing KD methods in QAT struggle to balance TS and distillation losses. We address this with a learnable regularizer and extend QAT-KD with QAT-EKD for better performance in low-bit quantization. Beyond discriminative tasks, quantization with distillation has also been investigated for generative models: Q-VDiT (Feng et al., 2025) targets video-generation diffusion transformers, while Li and Du (2025) propose cross-timestep error correction for quantized diffusion models. These works highlight the growing breadth of QAT-KD, though our focus is on small quantized models for classification, OD, and LLM compression.

3. Methodology

3.1. Problem Formulation

KD (Hinton et al., 2015) transfers knowledge from a high-capacity teacher network to a lightweight student by matching their output distributions. Let $\mathcal{D}s = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ be a dataset of inputs $\mathbf{x}_i \in \mathbb{R}^d$ and task targets $y_i \in \mathcal{Y}$. The student is a parametric function $f(\mathbf{x}; \theta)$, and the teacher is $f_T(\mathbf{x}; \theta_T)$ with frozen parameters θ_T. We denote by $\mathcal{L}_{\text{task}}$ any differentiable loss appropriate for the downstream task:

$$\mathcal{L}_{\text{task}} = \frac{1}{N} \sum_{i=1}^{N} \ell\big(f(\mathbf{x}_i; \theta),\, y_i\big), \tag{1}$$



(a) (b)

Figure 2: The proposed method learns to balance task-specific and distillation losses during QAT using a learnable regularizer. It employs adaptive components TSReg and KDReg to modulate gradients for stable and effective training. (a) QAT-KD framework with GoR. (b) Details of the proposed GoR.

where $\ell(\cdot,\cdot)$ represents the task-specific loss function.[1] KD is realized through a loss function that encourages the student model to mimic the teacher's predictions:

$$\mathcal{L}_{\text{KD}} = \mathcal{D}(f(\mathbf{x};\theta), f_T(\mathbf{x};\theta_T)) \tag{2}$$

where $\mathcal{D}(\cdot,\cdot)$ represents a distance measure between teacher and student outputs.[2] A conventional KD training step minimizes the weighted sum with a hand-tuned scalar $\alpha > 0$ balancing fidelity to ground-truth targets and mimicry of the teacher.

$$\mathcal{L} = (1-\alpha)\mathcal{L}_{\text{task}} \; + \; \alpha\,\mathcal{L}_{\text{KD}} \tag{3}$$

To deploy on edge hardware, we need to perform QAT, the student's full-precision parameters θ are mapped to low-precision $\theta_q = \text{Quantize}(\theta)$. During QAT the forward pass uses θ_q (simulating on-device behaviour) while gradients flow through the straight-through estimator. Replacing θ with θ_q in Eq. (3) yields

$$\mathcal{L}_{\text{QAT}} = (1-\alpha)\mathcal{L}_{\text{task}}\big(f(\mathbf{x};\theta_q), y\big) \; + \; \alpha\,\mathcal{L}_{\text{KD}}\big(f(\mathbf{x};\theta_q), f_T(\mathbf{x};\theta_T)\big). \tag{4}$$

Quantization introduces non-differentiable noise in the forward pass, whereas KD tries to smooth the student's prediction space. The fixed coefficient α in Eq. (4) is therefore notoriously fragile: too small, and the student overfits quantization artifacts; too large and task gradients vanish, causing convergence to stall. The proposed GoR introduced next replaces this static trade-off with two trainable, mutually balancing scalars that adapt online to the loss landscape.

[1] Common instantiations include cross-entropy for classification, smooth-L_1 for regression, focal loss for OD, etc. The generic formulation accommodates any differentiable task-specific objective.

[2] For classification tasks, $\mathcal{L}_{\text{KD}}$ is typically implemented as the Kullback–Leibler divergence between temperature-scaled softmax outputs. For regression and other non-classification tasks, appropriate distance measures such as mean squared error or cosine similarity are commonly used.

3.2. Game of Regularizers

Motivated by recent observations that fixed loss weighting α in Eq. (4), we introduce the GoR, sketched in Fig. 2(b). GoR replaces the static trade-off with two learnable scalars: α_{task} attached to the task loss $\mathcal{L}_{\text{task}}$ of Eq. (1) and α_{KD} attached to the distillation loss $\mathcal{L}_{\text{KD}}$ of Eq. (2). The two scalars interact antagonistically. Thus, neither can dominate the optimization. We denote the positive regularizers as $\alpha_{\text{task}}, \alpha_{\text{KD}} \in \mathbb{R}_{>0}$. The GoR objective:

$$\mathcal{L}_{\text{GoR}} \;=\; \frac{\alpha_{\text{task}}}{\alpha_{\text{KD}}}\,\mathcal{L}_{\text{task}} \;+\; \frac{\alpha_{\text{KD}}}{\alpha_{\text{task}}}\,\mathcal{L}_{\text{KD}}. \tag{5}$$

Conventional reweighting strategies such as fixed coefficients, gradient-norm balancing, or uncertainty-based weighting are not directly transferable to the quantized setting. In QAT, each loss gradient can be expressed as $\nabla_\theta \mathcal{L}_i(\theta_q) = g_i(\theta) + \xi_i(\theta)$, where $g_i(\theta)$ denotes the clean gradient and $\xi_i(\theta)$ is a parameter-dependent, biased, and heteroskedastic perturbation induced by quantization. The scale and distribution of ξ_i differ across the task and distillation losses, distorting both gradient magnitudes and loss statistics. Since existing weighting schemes rely on such statistics, these distortions often cause miscalibration or even invert the intended weighting effect (Zhao and Zhao, 2024).

The key insight is that GoR formulation creates a self-regulating system: as one scalar increases to emphasize its loss, it simultaneously reduces the influence of the other loss, preventing either from completely dominating the optimization process. The scalars are optimized along with other model parameters jointly:

$$\theta \leftarrow \theta - \eta_\theta\, \nabla_\theta \mathcal{L}_{\text{GoR}}, \tag{6}$$

$$\alpha_{\text{task}} \leftarrow \alpha_{\text{task}} - \eta_\alpha\, \nabla_{\alpha_{\text{task}}} \mathcal{L}_{\text{GoR}}, \tag{7}$$

$$\alpha_{\text{KD}} \leftarrow \alpha_{\text{KD}} - \eta_\alpha\, \nabla_{\alpha_{\text{KD}}} \mathcal{L}_{\text{GoR}}. \tag{8}$$

Here, η_θ and η_α are the learning rates for parameters and regularizers, respectively, and $\nabla_x \mathcal{L}_{\text{GoR}} := \partial \mathcal{L}_{\text{GoR}} / \partial x$ denotes the gradient of $\mathcal{L}_{\text{GoR}}$ with respect to its argument x. After each update we clip the scalars to $[10^{-4}, +\infty)$ to guarantee $\alpha_{\text{task}}, \alpha_{\text{KD}} > 0$. Differentiating Eq.(5) with respect to each loss term yields:

$$\frac{\partial \mathcal{L}_{\text{gor}}}{\partial \mathcal{L}_{\text{task}}} = \frac{\alpha_{\text{task}}}{\alpha_{\text{KD}}}, \qquad\qquad \frac{\partial \mathcal{L}_{\text{gor}}}{\partial \mathcal{L}_{\text{KD}}} = \frac{\alpha_{\text{KD}}}{\alpha_{\text{task}}}. \tag{9}$$

While Eq. (5) mathematically depends on the ratio of α_{task} and α_{KD}, the two-parameter formulation enables unique optimization dynamics through bidirectional competition. Each scalar experiences self-reinforcement and mutual inhibition as shown in (9), creating a game-theoretic equilibrium impossible with single-parameter alternatives like $\mathcal{L} = (1-\beta)\mathcal{L}_{\text{task}} + \beta\mathcal{L}_{\text{KD}}$. Table 4 empirically validates this design choice. Analyzing the partial derivatives of $\mathcal{L}_{\text{GoR}}$ with respect to the scalars makes it clearer:

$$\frac{\partial \mathcal{L}_{\text{GoR}}}{\partial \alpha_{\text{task}}} = \frac{1}{\alpha_{\text{KD}}}\,\mathcal{L}_{\text{task}} - \frac{\alpha_{\text{KD}}}{\alpha_{\text{task}}^2}\,\mathcal{L}_{\text{KD}}, \qquad \frac{\partial \mathcal{L}_{\text{GoR}}}{\partial \alpha_{\text{KD}}} = \frac{1}{\alpha_{\text{task}}}\,\mathcal{L}_{\text{KD}} - \frac{\alpha_{\text{task}}}{\alpha_{\text{KD}}^2}\,\mathcal{L}_{\text{task}}. \tag{10}$$

Setting these to zero yields the equilibrium

$$\alpha_{\text{task}}^2\,\mathcal{L}_{\text{task}} \;=\; \alpha_{\text{KD}}^2\,\mathcal{L}_{\text{KD}}, \tag{11}$$

Unlike prior approaches, the coupled formulation in Eq. (5) learns α_{task} and α_{KD} jointly with θ, such that the scalars adapt directly to the same biased gradients encountered in QAT-KD. The antagonistic balance constraint in Eq. (11) therefore provides stable optimization despite quantization-induced perturbations. Thus, the two losses are power-balanced, allowing each to retain its influence and jointly guide the training process. Equations (10) reveal the underlying game-theoretic mechanism: each scalar experiences both self-reinforcement (positive terms proportional to its associated loss) and mutual inhibition (negative terms that grow with the competing loss). This creates a natural equilibrium where the scalars dynamically balance to prevent either loss from dominating optimization. Notably, GoR introduces just two trainable scalars, $\mathcal{O}(1)$ parameters, and negligible FLOPs relative to the backbone.

3.3. Towards QAT–KD-GoR Framework

We propose a GoR-driven QAT–KD framework that dynamically balances task-specific and distillation objectives under quantization constraints. Notably, GoR is modular and can be seamlessly integrated into any existing KD method without requiring architectural modifications or changes to the underlying training pipeline. This makes it broadly applicable across diverse model architectures and tasks, enabling improved performance with minimal overhead. We begin by integrating the GoR weights from Eq. (5) into the QAT–KD objective of Eq. (4), yielding a dynamically balanced loss:

$$\mathcal{L}_{\text{GoR}}^{\text{QAT}} = \frac{\alpha_{\text{task}}}{\alpha_{\text{KD}}}\,\mathcal{L}_{\text{task}}\big(f(\mathbf{x};\theta_q), y\big) \;+\; \frac{\alpha_{\text{KD}}}{\alpha_{\text{task}}}\,\mathcal{L}_{\text{KD}}\big(f(\mathbf{x};\theta_q),\; f_T(\mathbf{x};\theta_T)\big) \tag{12}$$

where the learnable scalars $\alpha_{\text{task}}, \alpha_{\text{KD}} > 0$ dynamically balance task supervision and knowledge distillation during training. To further enhance generalization, we extend knowledge distillation to an ensemble of n teachers $\{T_1, \dots, T_n\}$. Let $\mathbf{z}_{T_i} = f_{T_i}(\mathbf{x}; \theta_{T_i}) \in \mathbb{R}^C$ denote the logits of teacher T_i. We form an ensemble teacher by averaging logits:

$$\mathbf{z}_{\text{ens}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{z}_{T_i}, \tag{13}$$

which is computationally free at inference and empirically competitive with more complex aggregation methods. The student model is then trained to minimize the discrepancy between its predictions and the ensemble teacher's predictions using a distance measure $\mathcal{D}(\cdot,\cdot)$, adapted for the quantized domain. The generalized EKD loss function can be represented by updating Eq. (2) as follows:

$$\mathcal{L}_{\text{EKD}} = \mathcal{D}\left(f(\mathbf{x};\theta_q), \mathbf{z}_{\text{ens}}\right) \tag{14}$$

where $\theta_q = \mathrm{Q}(\theta)$ represents the quantized student parameters. The loss function can be adjusted for various specific tasks by selecting an appropriate distance measure $\mathcal{D}(\cdot,\cdot)$. Replacing $\mathcal{L}_{\text{KD}}$ in Eq. (12) with $\mathcal{L}_{\text{EKD}}$ from Eq. (14) yields the unified objective:

$$\mathcal{L}_{\text{GoR+EKD}}^{\text{QAT}} = \frac{\alpha_{\text{task}}}{\alpha_{\text{KD}}}\,\mathcal{L}_{\text{task}}\big(f(\mathbf{x};\theta_q), y\big) \;+\; \frac{\alpha_{\text{KD}}}{\alpha_{\text{task}}}\,\mathcal{L}_{\text{EKD}} \tag{15}$$

When $n = 1$, $\mathcal{L}_{\text{EKD}}$ reduces to the conventional KD loss in Eq. (2), and Eq. (15) recovers Eq. (12). Thus, our formulation generalizes classical knowledge distillation to richer teacher ensembles with a learnable weighting scheme.

Algorithm 1 Training QAT-KD with Go with optional Ensemble KD

```
Input: dataset $\mathcal{D}$, pre-trained teacher(s) $\{f_{T_j}\}_{j=1}^{n}$, initial student weights $\theta$, quantization
       function $\mathrm{Q}(\cdot)$, temperature $\tau$, learning rates $\eta_\theta, \eta_\alpha$
// Initialize regularizers
$\alpha_{\text{task}} \leftarrow 1$, $\alpha_{\text{KD}} \leftarrow 1$   foreach mini-batch $\{(\mathbf{x}_i, y_i)\} \subset \mathcal{D}$ do
    $\theta_q \leftarrow \mathrm{Q}(\theta)$                                      // fake quantization
    $\mathcal{L}_{\text{task}} \leftarrow \ell\big(f(\mathbf{x}; \theta_q), y\big)$                          // Compute task loss
                                          // Compute knowledge distillation loss
    if $n > 1$ then
    $\mathbf{z}_{T_i} \leftarrow f_{T_i}(\mathbf{x}; \theta_{T_i})$ for $i = 1, \ldots, n$   $\mathbf{z}_{\text{ens}} \leftarrow \frac{1}{n}\sum_{i=1}^{n} \mathbf{z}_{T_i}$      // ensemble logits
    $\mathcal{L}_{\text{KD}} \leftarrow \mathcal{D}(f(\mathbf{x}; \theta_q), \mathbf{z}_{\text{ens}})$                         // Ensemble KD loss
end
else
        $\mathcal{L}_{\text{KD}} \leftarrow \mathcal{D}(f(\mathbf{x}; \theta_q), f_T(\mathbf{x}; \theta_T))$          // Standard single-teacher KD
end
$\mathcal{L}_{\text{GoR}} \leftarrow \frac{\alpha_{\text{task}}}{\alpha_{\text{KD}}} \mathcal{L}_{\text{task}} + \frac{\alpha_{\text{KD}}}{\alpha_{\text{task}}} \mathcal{L}_{\text{KD}}$                  // Game of Regularizers objective
$\{\theta_q, \alpha_{\text{task}}, \alpha_{\text{KD}}\} \leftarrow$ update via equations (6), (7), (8)          // Update parameters
$\alpha_{\text{task}} \leftarrow \max(\alpha_{\text{task}}, 10^{-4})$   $\alpha_{\text{KD}} \leftarrow \max(\alpha_{\text{KD}}, 10^{-4})$          // Clip regularizers
Output: quantized student $\theta_q^\star$ with learned regularizers $\alpha_{\text{task}}^\star, \alpha_{\text{KD}}^\star$
```

Implementation and training. We implement our framework using NVIDIA's open-source PyTorch-quantization library with fake-quantizer modules to emulate low-bit arithmetic. In our setup, only the student model is quantized during training. Floating-point tensors are scaled by $s = \frac{x_{\max} - x_{\min}}{2^n - 1}$, rounded, clipped, and dequantized as $x_q = s \cdot q + x_{\min}$, with gradients approximated via the straight-through estimator. KD is applied between the quantized student and the full-precision teacher (please refer to Fig. 2 (a)), using the KL divergence on temperature-scaled logits and optionally feature-based losses, depending on the distillation method. This allows the student to learn under realistic quantization noise while benefiting from both high-level and intermediate supervision. Algorithm 1 illustrates the training details of the proposed QAT-KD-GoR as a framework.

4. Experiments

4.1. Comparison with State-of-the-art Methods

We first demonstrate the practicality of the proposed GoR by integrating it into both logit-based and feature-based KD methods under QAT. We evaluate its effectiveness across diverse tasks, including image classification, OD, and LLM, under standard low-bit settings (e.g., 8-bit). Additionally, we explore more aggressive quantization (e.g., 4-bit) where applicable, excluding scenarios where it leads to severe performance degradation (e.g., OD).

4.1.1. Classification

The proposed GoR is a versatile regularization term that integrates seamlessly with existing KD methods. We evaluated its effectiveness by incorporating GoR into WSLD (Zhou et al., 2021), QFD (Zhu et al., 2023), and SQKD (Zhao and Zhao, 2024) for MobileNetV2 and ResNet18 students with a ResNet50 teacher on ImageNet at 8-bit and 4-bit precision. Contrary to SQKD's claim that KD loss alone suffices, adding GoR consistently improves performance (Table 1). For 8-bit quantization, GoR yields modest gains: +0.14% on MobileNetV2 (71.65% → 71.79%) and +0.11% on ResNet18 (69.64% → 69.75%). At 4-bit, where quantization challenges are severe, improvements are more pronounced: QAT-KD gains +3.28% on MobileNetV2 (55.72% → 59.01%) and +0.78% on ResNet18, while QFD achieves +3.37% on MobileNetV2. These results highlight GoR's effectiveness, particularly in low-precision scenarios.

Table 1: Comparison of QAT-KD methods on classification with and without GoR for MobileNetV2 and ResNet18.

Method	**ResNet50 → MobileNetV2**				**ResNet50 → ResNet18**			
	8/8		**4/4**		**8/8**		**4/4**	
	w/o GoR	GoR	w/o GoR	GoR	w/o GoR	GoR	w/o GoR	GoR
PTQ	71.08	–	0.70	–	69.39	–	0.35	–
QAT	71.36	–	43.82	–	69.56	–	60.30	–
QAT + KD (Hinton et al., 2015)	71.65	**71.79** (+0.14)	55.72	**59.01** (+3.28)	69.64	**69.75** (+0.11)	60.80	**61.57** (+0.78)
WSLD Zhou et al. (2021)	71.60	**71.71** (+0.11)	47.91	**48.35** (+0.45)	69.65	**69.71** (+0.06)	61.91	**62.09** (+0.18)
QFD Zhu et al. (2023)	71.73	**71.76** (+0.03)	56.00	**59.38** (+3.37)	69.73	**69.86** (+0.13)	60.95	**61.18** (+0.23)
SQKD Zhao and Zhao (2024)	71.43	**71.54** (+0.11)	47.13	**48.84** (+1.71)	69.56	**69.75** (+0.21)	60.94	**61.16** (+0.22)

4.1.2. LLM

We evaluated our GoR regularizer on LLMs by transferring knowledge from Qwen 2.5 3B (teacher) to Qwen 2.5 0.5B (student) (et al., 2025) under 8-bit and 4-bit quantization. We used the FineTome-100k dataset (mlabonne, 2025), a 100,000-example subset of arcee-ai/The-Tome (arcee ai, 2025), as our evaluation benchmark. Table 2 shows performance metrics for Qwen2.5 compression (3B → 0.5B) across quantization methods. PTQ exhibited the highest perplexity and lowest downstream metrics, especially at 4-bit. QAT improved upon PTQ at both precisions, while logit distillation (KD) (Hinton et al., 2015; Boizard et al., 2024) further enhanced performance, particularly at 8-bit (perplexity: 5.56). Our proposed GoR Logit Distillation achieved superior results across all metrics, with lowest perplexity (4.89 at 8-bit, 6.27 at 4-bit) and highest BLEU and BertScore, confirming the effectiveness of combining advanced logit distillation with QAT for low-precision quantization.

Table 2: Enhanced quantization results for LLM (Qwen2.5 3B → Qwen2.5 0.5B) at 8-bit and 4-bit precision, showing improvements.

Method	**8/8**					**4/4**				
	Perplexity ↓	MMLU ↑	Hellaswag ↑	BLEU ↑	BertScore ↑	Perplexity ↓	MMLU ↑	Hellaswag ↑	BLEU ↑	BertScore ↑
PTQ	7.74	23.15	30.15	7.94	77.6	9.56	22.87	28.70	7.53	75.01
QAT	6.57	26.46	31.33	8.56	83.17	6.85	22.95	29.74	8.02	77.84
QAT + KD (Boizard et al., 2024)	5.55	27.87	**33.14**	9.96	83.50	8.50	23.41	30.81	8.31	79.12
QAT + KD + GoR (Ours)	**4.89**	**28.38**	32.81	**12.73**	**85.96**	**6.27**	**23.94**	**31.52**	**8.84**	**83.15**
	(-0.66)	(+0.51)	(-0.33)	(+2.77)	(+2.46)	(-0.58)	(+0.53)	(+0.71)	(+0.53)	(+4.03)

4.1.3. Object Detection

We extended GoR beyond classification to OD by employing YOLOX-Small (Ge et al., 2021) as the student and YOLOX-Medium (Ge et al., 2021) as the teacher for QAT-KD, using the COCO dataset (Lin et al., 2014) for training and validation. We evaluated state-of-the-art KD for OD methods like CWD (Shu et al., 2021) and MGD (Yang et al., 2022) alongside baseline KD to benchmark performance. As shown in Table 3, GoR consistently improved QAT-KD performance, boosting mean average precision (mAP) at IoU 0.5 from 57.22 to 58.03 (+0.81) for CWD and from 57.68 to 59.2 (+1.52) for MGD. For mAP at IoU 0.5:0.95, GoR achieved improvements from 39.05 to 39.35 (+0.30) for CWD and from 39.25 to 39.48 (+0.23) for MGD. These results highlight GoR's effectiveness in enhancing OD under quantized settings, demonstrating its robustness across different tasks and architectures.

Table 3: Quantization performance comparison for YOLOX-M → YOLOX-S detection with and without GoR.

Method	**w/o GoR**		**GoR**	
	AP50	**AP50:95**	**AP50**	**AP50:95**
PTQ	56.81	38.75	–	–
QAT	57.23	39.05	–	–
CWD (Shu et al., 2021)	57.22	39.05	**58.03 (+0.81)**	**39.35 (+0.30)**
MGD (Yang et al., 2022)	57.68	39.25	**59.20 (+1.52)**	**39.48 (+0.23)**

4.2. Analysis of GoR

We evaluated learnable regularization against static weighting in QAT-KD ($\mathcal{L} = (1 - \alpha)\mathcal{L}_{\text{task}} + \alpha\mathcal{L}_{\text{KD}}$) on ImageNet with MobileNetV2 (8-bit). As shown in Table 4, static weights achieved Top-1 accuracies of 71.46% ($\alpha = 1.0$), 71.52% ($\alpha = 0.2$), and 71.62% ($\alpha = 0.5$), while our GoR framework with learnable α_{task} and α_{KD} achieved 71.79%. Furthermore, Table 5 shows that GoR outperforms existing reweighting methods, such as dynamic knowledge distillation (DKD) (Li et al., 2021a), demonstrating its adaptability across various models, datasets, and quantization schemes without requiring manual tuning.

Table 4: Comparison of GoR with static regularizers on QAT-KD.

Method	**Weight (α)**	**Top-1 (%)**
Static weighting	1.0	71.46
	0.2	71.52
	0.5	71.62
GoR (Ours)	Learnable	**71.79**

Table 5: Performance comparison with reweighting method on ImageNet and CIFAR-100 with MobileNet-v2.

Method	**ImageNet**	**CIFAR-100**
QAT + KD	71.65	76.71
QAT + Reweighting (DKD) (Li et al., 2021a)	71.52 (-0.13)	76.71 (+0.0)
QAT + KD + GoR	**71.79** (+0.14)	**76.90** (+0.19)

Moreover, Figure 3 visualizes the training dynamics of the QAT-KD for MobileNet-V2. Figures 3(a) and 3(b) illustrate the CE and KD losses with counter-reduction, respectively, showing that our approach successfully stabilizes training by dynamically balancing the interactions between loss components. Conversely, Figure 3(c) highlights the scenario without counter-reduction, where a single learnable regularizer keeps getting larger before getting clipped, leading to excessive amplification of quantization-induced distortions, hindering convergence and reducing overall accuracy. Figure 3(d) highlights the QAT-KD method along with the integration of dynamic regularization through GoR, which accelerates faster convergence.

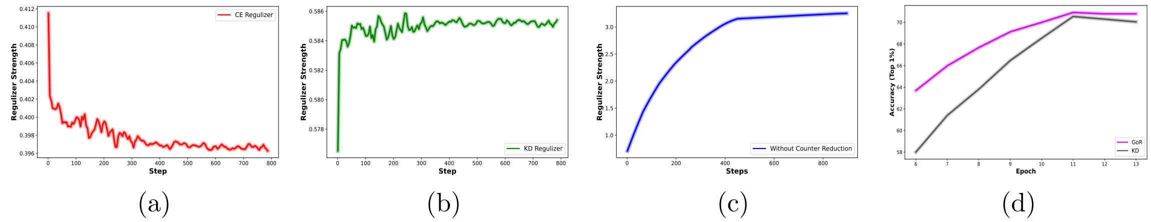


Figure 3: Effect of GoR on QAT-KD training stability for MobileNet-V2: (a-b) Balanced training with counter-reduction, (c) Instability without counter-reduction, and (d) convergence acceleration via dynamic regularization.

Figure 4 compares attention patterns across teacher, PTQ, QAT-KD, and GoR-QAT-KD models. Figure 4(a) shows GradCAM (Selvaraju et al., 2017) activation maps where GoR-QAT-KD closely matches the teacher's localization patterns. Figure 4(b) presents activation intensity histograms, with GoR-QAT-KD exhibiting similar distribution to the teacher. Figure 4(c) quantifies attention quality using peak-to-average ratio (measuring attention concentration) and Shannon entropy (capturing attention diversity). These complementary metrics illustrate that the SQM optimized with the proposed method preserves the teacher's balanced focus on salient regions. GoR-QAT-KD achieves values closest to those of the teacher for both metrics, while PTQ and QAT-KD exhibit larger deviations, confirming superior attention preservation.

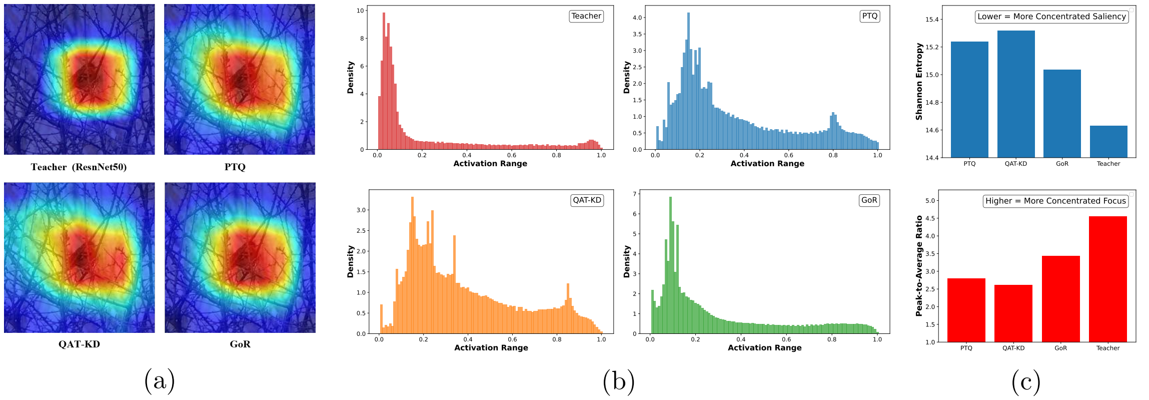


Figure 4: GoR-QAT-KD outperforms PTQ and QAT-KD in attention preservation: (a) Grad-CAM visualizations, (b) activation distributions, and (c) attention metrics.

4.3. EKD with GoR

We extend our evaluation of GoR beyond QAT-KD by exploring its integration with EKD under QAT, targeting edge-efficient models like MobileNetV2 that typically lack scalable variants and homogeneous teacher options. To address this, we leverage heterogeneous teacher ensembles (e.g., RegNet, ConvNeXt, Swin-T) within the EKD-GoR framework. As shown in Table 6, our method achieves 71.99% Top-1 accuracy, outperforming both single-teacher KD and homogeneous EKD setups—and even surpassing the full-precision baseline (71.87%).

We also compare the proposed EKD-GoR with existing multi-teacher methods under QAT. As shown in Table 7, EKD-GoR achieves the highest accuracy at both 8-bit (71.99%) and 4-bit (59.87%) precision, significantly outperforming recent approaches such as AMT (Liu et al., 2020), CMT (Wang et al., 2021), and CAMKD (Zhang et al., 2022). The substantial gain at 4 bits highlights the robustness and practicality of EKD-GoR for real-world low-bit deployments.

Table 6: Single vs multi-teacher performance comparison.

Method	**Teacher Type**	**Teacher**			**Student**
		Model	**Params (M)**	**Top-1**	**Top-1**
Baseline	N/A	–	–	–	71.87
PTQ	N/A	–	–	–	71.08
QAT	N/A	–	–	–	71.36
QAT + KD	Single	ResNet50	25.6	76.13	71.66 (+0.36)
QAT + KD	Single	RegNet	5.5	72.83	71.51 (+0.15)
QAT + KD	Single	ConvNeXt	28.6	82.52	71.43 (+0.08)
QAT + KD	Single	Swin-T	28.3	81.47	71.28 (-0.08)
QAT + EKD + GoR (Homogeneous)	Multi	ResNet50	25.6	76.13	71.82 (+0.46)
	Multi	ResNet101	44.50	77.37	
QAT + EKD + GoR (Proposed)	Multi	RegNet	5.5	72.83	71.99 (+0.63)
	Multi	ConvNeXt	28.6	82.52	
	Multi	Swin-T	28.3	81.47	

Table 7: Comparison between multi-teacher KD and the proposed EKD-GoR.

Method	**8/8**	**4/4**
AMT (Liu et al., 2020)	71.58	53.17
CMT (Wang et al., 2021)	71.71	55.75
CAMKD (Zhang et al., 2022)	71.33	55.46
EKD + GoR (Ours)	**71.99**	**59.87**

4.4. Real-world Edge Performance

We evaluated the proposed QAT-KD-GoR framework by deploying the optimized SQM on real-edge hardware, such as the Jetson Orin. To generalize beyond the MobileNet–ImageNet setup, we extended our study to include commonly used lightweight architectures such as VGG, ResNet18, and RegNetX-400MF (Radosavovic et al., 2020), as well as transformer-based models like Swin-T (Liu et al., 2021). We further examined its applicability to full-precision training on smaller datasets like CIFAR-100, using transfer learning with ImageNet-pretrained weights while maintaining consistency. Table 8 reports accuracy and FPS across ImageNet and CIFAR-100 under various Jetson Orin power settings (MaxN, 15W, 30W, 50W). The QAT-KD-GoR framework consistently outperformed both baseline and standard QAT+KD. For example, ResNet50 as a teacher yielded a 0.47% gain for ResNet18 on ImageNet and improved FPS; RegNet improved accuracy by 0.24% with better performance at lower power. Swin-T achieved a 0.26% gain on ImageNet, along with notable FPS improvements, especially at 15W. On CIFAR-100, Swin-T showed the largest gain of 1.67%. These results demonstrate the robustness and deployment readiness of the QAT-KD-GoR framework across architectures, datasets, and power-constrained environments. Notably, the proposed method can compress SQMs to achieve up to a 242% increase in inference speed compared to the base model, without any loss in accuracy—and in some cases, even achieving better accuracy than the full-precision baseline.

Table 8: Quantization accuracy (Top-1 %) with FPS across ImageNet and CIFAR100 datasets. Accuracy gains over baseline/QAT+KD shown in parentheses.

Model	**ImageNet**				**CIFAR100**				**Speed on Jetson (FPS)**							
	Base	KD + GoR (FP32)	QAT+KD	Our (INT8)	Base	KD + GoR (FP32)	QAT+KD	Our (INT8)	MaxN FP32	MaxN INT8	15W FP32	15W INT8	30W FP32	30W INT8	50W FP32	50W INT8
Resnet18	69.76	**70.23** (+0.47)	69.64	**69.71** (+0.11)	80.02	**80.19** (+0.17)	79.89	**80.01** (+0.12)	1087	2447	167	623	381	1213	736	1856
VGG	71.59	**71.65** (+0.06)	71.52	**71.64** (+0.12)	78.95	**79.48** (+0.53)	78.93	**79.22** (+0.29)	166	568	30	121	60	229	121	414
MobileNet-v2	71.87	**72.16** (+0.29)	71.65	**71.79** (+0.14)	76.83	**77.07** (+0.24)	76.71	**76.90** (+0.19)	948	1103	182	295	338	461	578	717
RegNetX-400MF	72.83	**73.07** (+0.24)	72.79	**72.97** (+0.18)	81.73	**81.90** (+0.17)	81.57	**81.70** (+0.13)	1169	1805	196	491	456	823	733	1151
Swin-T	81.47	**81.73** (+0.26)	81.46	**81.59** (+0.13)	80.32	**81.99** (+1.67)	80.29	**80.50** (+0.21)	117	134	20	25	37	45	73	86

4.5. Discussion

We demonstrate the practicality of the GoR-driven QAT-KD framework for optimizing SQM in real-world applications. Our results show that a learnable regularizer with just two parameters can significantly enhance existing KD methods under QAT. Additionally, leveraging heterogeneous teachers via EKD-GoR narrows the performance gap with full-precision models and can even surpass baseline models under optimal conditions. QAT+KD+GoR is a versatile strategy that could be applicable to various vision tasks, including image-to-image translation, segmentation, detection, and LLMs. We plan to explore its broader applications in future work.

5. Conclusion

In this study, we propose a novel learnable regularization strategy to optimize SQMs. We demonstrate that only two learnable parameters can effectively address the well-known QAT-KD loss balancing issues. This results in significantly reduced quantization error, enabling efficient knowledge transfer from large models to SQMs. This strategy boosts the performance of existing methods across various AI tasks and domains under aggressive quantization. Furthermore, by incorporating EKD with multiple heterogeneous teacher models, our framework further refines the student's distribution and can even surpass full-precision models under optimal conditions. We plan to explore the broader applicability of the proposed EKD-GoR framework to diverse vision and AI tasks in future work.

References

SM A Sharif, Azamat Myrzabekov, Nodirkhuja Khudjaev, Roman Tsoy, Seongwan Kim, and Jaeho Lee. Learning optimized low-light image enhancement for edge vision tasks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6373–6383, 2024.

arcee ai. The-tome. https://huggingface.co/datasets/arcee-ai/The-Tome, 2025. Accessed: 2025-06-18.

Nicolas Boizard, Kevin El Haddad, Céline Hudelot, and Pierre Colombo. Towards cross-tokenizer distillation: the universal logit distillation loss for llms. arXiv preprint arXiv:2402.12030, 2024.

Jihoon Boo et al. Speq: Stochastic precision teacher for quantization-aware distillation. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2021a.

Yoonho Boo, Sungho Shin, Jungwook Choi, and Wonyong Sung. Stochastic precision ensemble: self-knowledge distillation for quantized deep neural networks. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, pages 6794–6802, 2021b.

Matthieu Courbariaux et al. Binarized neural networks: Training deep neural networks with weights and activations constrained to +1 or -1. In Advances in Neural Information Processing Systems (NeurIPS), 2016.

An Yang et al. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Weilun Feng, Chuanguang Yang, Haotong Qin, Xiangqi Li, Yu Wang, Zhulin An, Libo Huang, Boyu Diao, Zixiang Zhao, Yongjun Xu, et al. Q-vdit: Towards accurate quantization and distillation of video-generation diffusion transformers. arXiv preprint arXiv:2505.22167, 2025.

Zhi Ge, Wengang Liu, Yang Li, Jian Sun, Yawei Zhang, and Hongdong Zhang. Yolox: Exceeding yolo series in 2021. arXiv preprint arXiv:2107.08430, 2021.

Ruihao Gong et al. Differentiable soft quantization: Bridging full-precision and low-bit neural networks. In International Conference on Computer Vision (ICCV), 2019.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531, 2015.

Sangil Jung et al. Learning to quantize deep networks by optimizing quantization intervals with task loss. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

Eunsu Kim, Sungjoon Choi, and Seungyong Kim. Qkd: Knowledge distillation via quantization and denoising. IEEE Transactions on Neural Networks and Learning Systems, 30 (3):831–844, 2019.

Raghuraman Krishnamoorthi. Quantizing deep convolutional networks for efficient inference: A whitepaper. arxiv 2018. arXiv preprint arXiv:1806.08342, 1806.

Lei Li, Yankai Lin, Shuhuai Ren, Peng Li, Jie Zhou, and Xu Sun. Dynamic knowledge distillation for pre-trained language models. arXiv preprint arXiv:2109.11295, 2021a.

Yanxi Li and Chengbin Du. Optimizing quantized diffusion models via distillation with cross-timestep error correction. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 39, pages 18530–18538, 2025.

Zhen Dong Li et al. Mqbench: Towards reproducible and deployable model quantization benchmark. In Advances in Neural Information Processing Systems (NeurIPS), 2021b.

Zhuangwei Li et al. Apot: Adaptive rounding for post-training quantization. In Advances in Neural Information Processing Systems (NeurIPS), 2019.

Lightly.ai. Knowledge distillation trends. Lightly Blog, 2025. https://www.lightly.ai/blog/knowledge-distillation.

Tsung-Yi Lin, Jun Ma, Peter N Belhumeur, Dinesh Batra, Daniel Neumann, Jia Deng, Lei Xie, Zhicheng Hu, Shih-Fu Li, and Li Fei-Fei. Microsoft coco: Common objects in context. In European conference on computer vision (ECCV), pages 740–755, 2014.

Yuang Liu, Wei Zhang, and Jun Wang. Adaptive multi-teacher multi-level knowledge distillation. Neurocomputing, 415:106–113, 2020.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF international conference on computer vision, pages 10012–10022, 2021.

Abhishek Mishra, Eriko Nurvitadhi, Jeff Cook, and Debbie Marr. Apprentice: Using knowledge distillation techniques to improve low-precision network accuracy. In International Conference on Learning Representations (ICLR), 2018.

mlabonne. Finetome-100k: A diverse dataset for fine-tuning large language models. Dataset Repository, 2025. URL https://huggingface.co/datasets/mlabonne/FineTome-100k.

Cuong Pham, Tuan Hoang, and Thanh-Toan Do. Collaborative multi-teacher knowledge distillation for learning low bit-width deep neural networks. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 6435–6443, 2023a.

Quang Hieu Pham et al. Cmt-kd: Collaborative mutual teaching for quantized knowledge distillation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023b.

Ibraheem Radosavovic, Xiyang Zhang, Ross Girshick, and Kaiming He. Designing network design spaces. In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pages 10428–10437, 2020.

Christoffer Bøgelund Rasmussen, Aske Rasch Lejbølle, Kamal Nasrollahi, and Thomas B Moeslund. Evaluation of edge platforms for deep learning in computer vision. In International Conference on Pattern Recognition, pages 523–537. Springer, 2021.

Mohammad Rastegari et al. Xnor-net: Imagenet classification using binary convolutional neural networks. In European Conference on Computer Vision (ECCV), 2016.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE international conference on computer vision, pages 618–626, 2017.

Changyong Shu, Yifan Liu, Jianfei Gao, Zheng Yan, and Chunhua Shen. Channel-wise knowledge distillation for dense prediction, 2021. URL https://arxiv.org/abs/2011.13256.

Jindong Wang, Yiqiang Chen, Han Yu, et al. Collaborative multi-teacher knowledge distillation. IEEE Transactions on Neural Networks and Learning Systems, 32(8):3583–3594, 2021.

Zhendong Yang, Zhe Li, Mingqi Shao, Dachuan Shi, Zehuan Yuan, and Chun Yuan. Masked generative distillation, 2022. URL https://arxiv.org/abs/2205.01529.

Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. arXiv preprint arXiv:1612.03928, 2016.

Hailin Zhang, Defang Chen, and Can Wang. Confidence-aware multi-teacher knowledge distillation. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 4498–4502. IEEE, 2022.

Kaiqi Zhao and Ming Zhao. Self-supervised quantization-aware knowledge distillation. In Proceedings of the 27th International Conference on Artificial Intelligence and Statistics, volume 238 of Proceedings of Machine Learning Research, pages 4375–4383. PMLR, 2024. URL https://proceedings.mlr.press/v238/zhao24d.html.

X. Zhao et al. Intermediate representation transfer for knowledge distillation. In Proceedings of the AAAI Conference on Artificial Intelligence, 2023.

Helong Zhou, Liangchen Song, Jiajie Chen, Ye Zhou, Guoli Wang, Junsong Yuan, and Qian Zhang. Rethinking soft labels for knowledge distillation: A bias-variance tradeoff perspective. arXiv preprint arXiv:2102.00650, 2021.

Shuchang Zhou, Yuxin Wu, Zekun Ni, Xinyu Zhou, He Wen, and Yuheng Zou. Dorefa-net: Training low bitwidth convolutional neural networks with low bitwidth gradients. arXiv preprint arXiv:1606.06160, 2016.

Ke Zhu, Yin-Yin He, and Jianxin Wu. Quantized feature distillation for network quantization. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 37, pages 11452–11460, 2023.

Bohan Zhuang, Chunhua Shen, Mingkui Tan, Lingqiao Liu, and Ian Reid. Towards effective low-bitwidth convolutional neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 7920–7928, 2018.