Fast-Converging and Architecture-Agnostic 6-Bit Face Recognition via Gradient Coordination and Refined Data

Kim Thien Hoang[1] and Chien-Quang Le[1,*]
[1]Department of Information Technology, University of Sciences, Hue University, Hue, Vietnam
Email: {22T1020444, lqchien}@husc.edu.vn

***Abstract*—Deploying face recognition models on resource-constrained Edge AI devices requires aggressive model compression, where low-bit quantization is a promising solution. However, extreme 6-bit (Q6) quantization often distorts the embedding space, severely degrading recognition performance. In this paper, we propose an architecture-agnostic gradient-coordinated quantization framework that stabilizes training by aligning the student network's optimization trajectory with a full-precision teacher. Specifically, we introduce a β-weighted Mean Squared Error (β-MSE) alignment objective to amplify guidance signals, combined with a three-phase training strategy to stabilize dynamic ranges and optimization. Departing from the conventional reliance on large-scale synthetic datasets, we adopt a refined dataset strategy using approximately 53,500 real-world images enriched with high-quality samples covering diverse pose and age variations. Experimental results on standard benchmarks including LFW, CFP-FP, AgeDB-30, CALFW and CPLFW show that our method achieves 57.4x faster convergence (3,135 vs. 180,000 iterations) compared to real-data-based state-of-the-art approaches like QuantFace when normalized to the same batch size. Our framework recovers nearly all performance loss, achieving a 102% recovery rate on AgeDB-30, and demonstrates robustness across multiple architectures, including iResNet-50 and MobileFaceNet.**

***Index Terms*—Face recognition, model quantization, low-bit quantization, gradient coordination, knowledge distillation, data-efficient learning, edge AI**

I. Introduction

Recent advances in deep face recognition have been largely driven by powerful deep architectures such as iResNet and margin-based loss functions like ArcFace [1] and CosFace [2], which enhance inter-class separability. However, the substantial computational complexity of these full-precision teacher models makes them unsuitable for resource-constrained edge devices. While aggressive 6-bit (Q6) quantization offers a promising solution for deployment efficiency [3], [4], it often leads to severe distortion in the embedding space, degrading recognition performance. This creates a fundamental conflict: maintaining the high discriminative accuracy of a full-precision teacher while satisfying the strict memory constraints of a Q6 student model.

To mitigate this, a common strategy is to leverage large-scale synthetic datasets, as seen in QuantFace [5], which relies on hundreds of thousands of generated images to recover performance loss. While effective, this paradigm introduces extremely high computational and data generation costs, limiting its practicality. Moreover, the reliance on synthetic data raises concerns regarding the domain gap between synthetic and real-world distributions.

In this work, we argue that such large-scale data dependency is unnecessary. Instead, we propose an efficient approach focusing on a relatively small, refined training dataset of approximately 53,500 real-world images from sources such as CASIA-WebFace [17], VGGFace2 [19], and IMDB-WIKI [18]. Specifically, we introduce a gradient-aligned objective based on a β-weighted mean squared error (β-MSE) loss combined with a three-stage training strategy. To provide a clear overview of the proposed method, we illustrate the full framework in Fig. 1. Experimental results confirm that our method achieves 102% recovery rate on AgeDB-30 and 57.4$\times$ faster convergence (3,135 vs. 180,000 iterations) than state-of-the-art approaches [5] when normalized to the same batch size.

The main contributions are: (1) A β-MSE objective to preserve embedding consistency; (2) A three-stage strategy to stabilize optimization; and (3) A demonstration of high recovery using a small refined dataset for edge AI environments.

II. Related Work

Recent advances in model compression have shown that quantization is an effective way to reduce memory usage and computational cost for edge deployment [3], [4]. However, applying aggressive low-bit quantization to face recognition (FR) is particularly challenging, as it tends to distort fine-grained geometric relationships in the embedding space and weaken inter-class margins originally established by margin-based losses like ArcFace [1] or CosFace [2].

A representative task-specific effort is *QuantFace*, which leverages large-scale synthetic datasets to recover performance loss in low-bit settings [5]. While effective, this approach relies heavily on extensive external resources, including a complex generative pipeline and significant training time for hundreds of thousands of generated images. Such heavy dependency on synthetic data scale limits its practicality in resource-constrained scenarios where data preparation budgets are tight. Furthermore, the inherent domain gap between synthetic and real-world distributions remains a concern for robust performance.

* Corresponding author

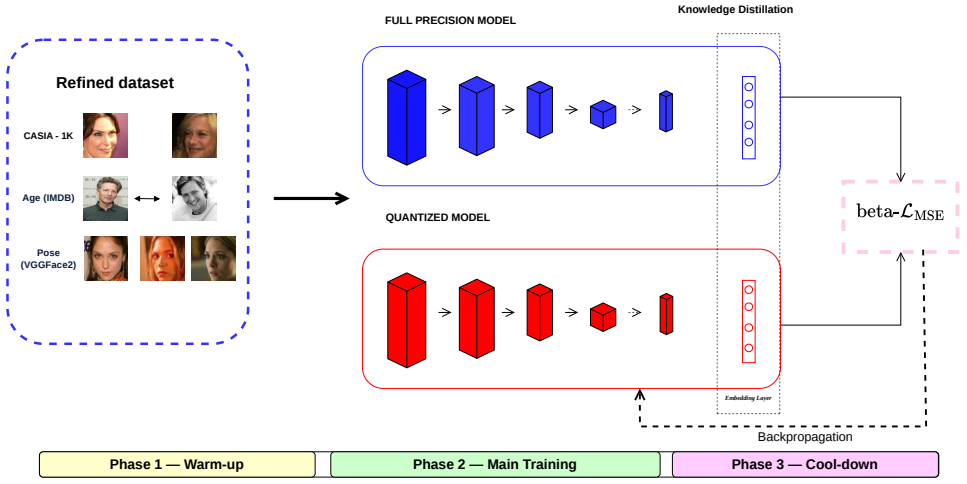

Fig. 1. Overview of the proposed framework. A full-precision teacher guides a 6-bit quantized student through β-MSE embedding alignment, a three-phase training schedule, and refined real-data subsets emphasizing pose and age variation.

To bridge the performance gap without massive retraining, knowledge distillation (KD) has become a standard paradigm, where a teacher network provides guidance through soft targets [7] or intermediate representation hints [8]. While KD typically focuses on matching final features, our work argues that coordinating the student's optimization trajectory through gradient coordination is equally critical for low-bit recovery.

Complementing this, recent research on data-efficient learning suggests that high recognition accuracy can be achieved without massive corpora. This is exemplified by the "baby learning" philosophy, where models learn effectively from small but highly informative datasets. Specifically, Tran *et al.* [11] demonstrated that vision transformers can achieve promising results under a baby-learning setting. Motivated by this, our framework departs from scaling synthetic data volume, focusing instead on a *refined dataset strategy* using approximately 53,500 real-world images from sources such as CASIA-WebFace [17], VGGFace2 [19], and IMDB-WIKI [18]. By combining informative real-data samples with gradient-coordinated supervision (β-MSE), we aim to achieve faster convergence and superior recovery compared to state-of-the-art synthetic-based approaches.

III. Proposed Methodology

This section presents the proposed framework for low-bit face recognition. The central idea is to quantize an iResNet-18 model to 6-bit precision while preserving the discriminative structure of the full-precision embedding space. To achieve this goal, we combine three elements: a Q6 quantization framework, a teacher-guided embedding alignment objective based on β-MSE, and a staged training strategy designed to stabilize optimization under severe precision reduction.

A. Low-bit Quantization Framework

Let $f_t(\cdot)$ denote the full-precision teacher network and $f_s(\cdot)$ denote the quantized student network. Both networks share the same iResNet-18 architecture and produce 512-dimensional face embeddings. The teacher remains in full precision and serves as a stable reference model, whereas the student is transformed into a low-bit network in which both weights and activations are represented with 6-bit precision. This setting is referred to as Q6 quantization.

For a real-valued tensor x, the quantization operation can be expressed as

$$Q_b(x) = \mathrm{clip}\left(\mathrm{round}\left(\frac{x}{s}\right) + z, q_{\min}, q_{\max}\right), \tag{1}$$

where b denotes the bit-width, s is the quantization scale, z is the zero-point, and $q_{\min}$ and $q_{\max}$ define the integer range associated with b. The dequantized representation is then given by

$$\hat{x} = s\left(Q_b(x) - z\right). \tag{2}$$

In the proposed framework, this mapping is applied throughout the student network so that convolutional weights and intermediate activations are both constrained to 6-bit representations.

Although quantization improves efficiency, it also introduces a mismatch between the full-precision and low-bit feature distributions. In face recognition, this mismatch is especially harmful because verification performance depends on preserving fine-grained relationships in the embedding space. Even a small perturbation can weaken intra-class compactness or inter-class separability. Therefore, the quantized student is not optimized independently; instead, it is guided to remain close to the embedding geometry of the full-precision teacher.

B. Gradient Coordination via β-MSE

For an input face image x, the teacher and student produce embeddings

$$\mathbf{t} = f_t(x), \qquad \mathbf{s} = f_s(x), \tag{3}$$

which are normalized to unit length:

$$\tilde{\mathbf{t}} = \frac{\mathbf{t}}{\|\mathbf{t}\|_2}, \qquad \tilde{\mathbf{s}} = \frac{\mathbf{s}}{\|\mathbf{s}\|_2}. \tag{4}$$

The alignment loss is defined as

$$\mathcal{L}_{\mathrm{GC}} = \beta \cdot \frac{1}{N}\sum_{i=1}^{N} \left\|\tilde{\mathbf{s}}_i - \tilde{\mathbf{t}}_i\right\|_2^2, \tag{5}$$

TABLE I
ACCURACY (Q6) COMPARISON ACROSS ARCHITECTURES AND BENCHMARKS.

Architecture	Model Type	Training Images	LFW	CFP-FP	AgeDB-30	CALFW	CPLFW
Accuracy (Q6)							
iresnet18	FP32 [5]	5.8M	99.617	93.671	96.687	95.533	89.183
	RealQuantFace [5]	5.8M	99.520	93.230	96.550	95.580	88.370
	SynQuantFace [5]	0.5M	99.550	**93.340**	96.620	95.320	**89.050**
	PTQ	–	99.500	92.671	96.633	95.283	87.733
	Ours (wo-3p)	0.053M	**99.633**	92.729	**96.817**	**95.600**	88.800
	Ours (w-3p)	0.053M	99.617 (β=50)	92.857 (β=300)	96.450 (β=200)	95.417 (β=1)	88.917 (β=1)
iresnet50	FP32 [5]	5.8M	99.800	95.957	97.983	96.083	92.217
	RealQuantFace [5]	5.8M	99.700	95.000	97.170	**95.870**	90.170
	SynQuantFace [5]	0.5M	99.680	**95.170**	97.430	95.700	90.380
	PTQ	–	99.683	91.557	96.083	95.133	87.017
	Ours (wo-3p)	0.053M	99.686	93.786	97.417	95.817	90.217
	Ours (w-3p)	0.053M	**99.750** (β=300)	92.857 (β=300)	**97.683** (β=200)	95.817 (β=200)	**91.767** (β=200)
mobilefacenet	FP32 [5]	5.8M	99.433	91.529	95.567	95.150	87.800
	RealQuantFace [5]	5.8M	98.870	87.690	93.030	93.300	84.570
	SynQuantFace [5]	0.5M	99.080	87.640	91.770	93.480	84.850
	PTQ	–	98.150	83.400	89.267	91.067	77.017
	Ours (wo-3p)	0.053M	98.883	87.314	92.450	93.317	83.967
	Ours (w-3p)	0.053M	**99.317** (β=300)	**90.514** (β=200)	**93.867** (β=200)	**93.650** (β=200)	**86.600** (β=300)

where N is the mini-batch size and β is a scaling coefficient.

This objective plays a central role in the proposed method. Rather than only requiring the student to produce similar final features, it continuously constrains the quantized model to follow the representation structure of the teacher during optimization. Because the student operates under low-bit noise, the raw mean squared error between normalized embeddings is typically small in magnitude and may not produce sufficiently strong corrective gradients. The coefficient β is therefore introduced to amplify the alignment signal and make the penalty for feature mismatch more influential during training.

A large value of β is important because it strengthens the attraction of the student toward the teacher embedding manifold. In practice, the quantized student is affected by discretization error at many layers, which can gradually shift its feature distribution away from the full-precision reference. If the alignment term is too weak, this deviation accumulates and leads to unstable or suboptimal convergence. By increasing the weight of the MSE term, the proposed loss forces the student to correct such deviations more aggressively. Consequently, the student is encouraged to preserve the discriminative geometry of the teacher even when operating under 6-bit constraints.

For this reason, we refer to the proposed objective as gradient coordination. The purpose is not only to reduce feature discrepancy at convergence, but also to coordinate the direction of parameter updates throughout training. In this way, the student is guided by a stronger and more stable optimization signal, which is particularly beneficial in low-bit face recognition where embedding distortion is a major source of performance degradation.

C. 3-Phase Training Recipe

In addition to the alignment objective, stable optimization is achieved through a three-phase training recipe. This design is motivated by the observation that quantized networks are highly sensitive to calibration and optimization noise, especially during the early stage of training when activation ranges are still unstable. As summarized in Fig. 1, the proposed schedule progresses from warm-up, to range stabilization, and finally to fine-tuning with learning-rate decay.

Phase 1: Warm-up. In the first phase, the student is gradually adapted to the quantized regime. During this stage, the quantization modules continue to collect activation statistics in order to estimate appropriate dynamic ranges. At the same time, training begins with a small learning rate and increases progressively to the base level. This warm-up process reduces the risk of abrupt optimization instability and allows the student to adjust to quantization error in a controlled manner.

Phase 2: Freeze Observers. Once the activation ranges have been sufficiently estimated, the observer modules are frozen. The purpose of this stage is to keep the quantization ranges fixed so that the student is optimized under a stable numerical configuration. Without this step, continuously changing ranges may introduce additional oscillation into the learning process. By fixing the dynamic range, the optimization becomes more consistent and the student can concentrate on aligning its embeddings with those of the teacher.

Phase 3: Fine-tuning. In the final phase, the model is refined under the fixed quantization setting. The learning rate is gradually reduced so that the student can make smaller and more precise adjustments around a good low-bit solution. This stage improves convergence quality and helps reduce the remaining discrepancy between the teacher and student embeddings. Since the optimization is performed on real training data rather than large-scale synthetic data, the final model remains consistent with the data-efficient objective of the proposed framework.

The overall learning-rate schedule is defined as

$$\eta_e = \begin{cases} \eta_{\min}^{\text{warm}} + \dfrac{e}{4}\left(\eta_0 - \eta_{\min}^{\text{warm}}\right), \\ \eta_0, \\ \eta_{\min}^{\cos} + \left(\eta_0 - \eta_{\min}^{\cos}\right)\dfrac{1+\cos\left(\pi(e-25)/5\right)}{2}, \end{cases} \tag{6}$$

where η_0 denotes the base learning rate, while $\eta_{\min}^{\text{warm}}$ and $\eta_{\min}^{\cos}$ denote the minimum learning rates used in the warm-up and

TABLE II
TAR@FAR= 10^{-4} COMPARISON ACROSS ARCHITECTURES AND BENCHMARKS.

Architecture	Model Type	Training Images	LFW	CFP-FP	AgeDB-30	CALFW	CPLFW
TAR@FAR=10^{-4}							
iresnet18	FP32 [5]	5.8M	99.133	77.686	84.967	85.000	50.733
	PTQ	–	99.033	71.429	84.000	83.300	**56.933**
	Ours (wo-3p)	0.053M	99.067	73.429	76.033	83.900	49.567
	Ours (w-3p)	0.053M	**99.233** (β=50)	**76.543** (β=300)	**87.900** (β=200)	**87.267** (β=1)	53.100 (β=1)
iresnet50	FP32 [5]	5.8M	99.600	88.914	92.900	90.500	53.533
	PTQ	–	98.500	67.343	76.400	82.833	0.100
	Ours (wo-3p)	0.053M	**99.433**	81.514	88.300	**90.433**	**56.233**
	Ours (w-3p)	0.053M	99.333 (β=300)	**85.657** (β=300)	**92.700** (β=200)	89.333 (β=200)	37.000 (β=200)
mobilefacenet	FP32 [5]	5.8M	98.167	65.857	74.500	84.700	9.600
	PTQ	–	91.600	25.286	24.600	49.600	0.067
	Ours (wo-3p)	0.053M	94.767	31.800	31.800	**76.000**	**15.133**
	Ours (w-3p)	0.053M	**95.900** (β=300)	**62.914** (β=200)	**44.400** (β=200)	73.200 (β=200)	4.033 (β=300)

fine-tuning stages, respectively.

Overall, the proposed methodology integrates low-bit quantization, strong teacher-guided embedding alignment, and staged optimization into a unified framework. Through this design, the student model is encouraged to retain the discriminative structure of the full-precision embedding space while remaining efficient enough for practical deployment on resource-constrained devices.

IV. Refined Dataset for Quantization

Our refined training set is constructed from three parts: CASIA-1k, a pose-focused subset, and an age-focused subset. CASIA-1k serves as the identity backbone and contains 48,458 images selected from CASIA-WebFace [17]. To explicitly strengthen robustness to viewpoint variation, we further build a Pose subset from VGGFace2 [19] by selecting 200 identities with large pose diversity and retaining 10 images per identity, resulting in 2,000 pose-oriented images. To improve tolerance to age-related appearance changes, we also build an Age subset from IMDB-WIKI [18] by selecting 300 identities and keeping 10 images per identity, resulting in 3,000 age-oriented images. As summarized in Fig. 1, these additional real-image subsets are not introduced to maximize data scale, but to inject the specific variations that are most useful for low-bit face recognition.

This selection strategy keeps the main identity structure from CASIA-1k while supplementing it with controlled pose and age variations. The role of CASIA-1k is to provide the student with a stable identity-rich training backbone, because its 48,458 images still preserve broad inter-class discrimination and sufficient intra-class variation for representation learning. On top of this base set, the Pose subset is intentionally limited to 200 identities with 10 images per identity so that each selected subject contributes a compact but diverse set of viewpoints. This design encourages the student to observe large pose changes within the same identity without overwhelming the training set with redundant samples. In parallel, the Age subset is constructed from 300 identities with 10 images per identity to expose the model to identity-preserving but age-related appearance changes, which are especially challenging in face verification. Such real variations are particularly valuable under 6-bit quantization because they enlarge intra-class differences and reveal the kinds of feature distortions that are easily amplified by quantization noise. As a result, the student is trained not only to preserve class-level discrimination, but also to remain stable under realistic pose and age changes that commonly cause performance degradation. Therefore, the refined dataset complements the proposed β-MSE objective and three-phase training schedule by improving robustness through informative real-world variation rather than data volume alone. The rationale for refining the dataset is twofold. First, 6-bit quantization limits weights and activations to only 64 levels, making face embeddings more sensitive to challenging variations such as pose and age. Therefore, we emphasize these difficult cases using a compact real-world dataset rather than large-scale redundant data. Second, with only 53,500 images (110× fewer than the 5.8M-image setting of QuantFace), the proposed dataset significantly reduces training cost while enabling fast convergence and efficient edge deployment.

V. Experimental Results

A. Experimental Setup

All experiments use the proposed Q6 training framework with stochastic gradient descent (SGD), momentum 0.9, weight decay 5×10^{-4}, and batch size 128. The optimization follows the three-phase schedule described earlier, and the embedding alignment objective is optimized with the proposed β-MSE loss. We evaluate the method on LFW [12], CFP-FP [13], AgeDB-30 [14], CALFW [15], and CPLFW [16] under two complementary metrics: verification accuracy and TAR@FAR= 10^{-4}.

B. Summary Results

Table I summarizes the Q6 accuracy results across three architectures, while Table II reports the corresponding TAR@FAR= 10^{-4} results. Table III further reports the recovery-rate ablation between CASIA-1k and CASIA-1k augmented with the proposed new data. The comparison includes the full-precision teacher (FP32), PTQ baselines, prior QuantFace variants when available, and the proposed method with and without the three-phase training recipe. Here, PTQ denotes post-training quantization, where the pretrained full-precision

TABLE III
RECOVERY-RATE ABLATION FOR CASIA-1K AND CASIA-1K + NEW DATA.

Training Data	LFW	CFP-FP	AgeDB-30	CALFW	CPLFW
CASIA-1k	0.999	0.949	1.007	1	0.21
CASIA-1k + New Data	1.0	0.985	1.008	1.03	1.05

model is directly quantized to Q6 without additional teacher-guided fine-tuning; it is used as the basic low-bit baseline for measuring the benefit of the proposed training strategy. The recovery-rate values in Table III measure how much of the teacher's TAR performance is retained by the Q6 student:

$$\textit{RecoveryRate} = \frac{TAR_{\text{Student\_Q6}}}{TAR_{\text{Teacher\_FP32}}} \tag{7}$$

The summary tables highlight three main observations. First, refined real-data training consistently improves over direct PTQ while using only 0.053M training images, and the ablation in Table III shows that adding the proposed pose- and age-focused data improves recovery on most benchmarks compared with CASIA-1k alone. Second, the three-phase schedule generally improves stability and gives the strongest overall results for the proposed method, especially on CFP-FP and AgeDB-30. Third, the method is architecture-agnostic: the same training principle benefits iResNet-18, iResNet-50, and MobileFaceNet, although the lightweight MobileFaceNet remains more sensitive under the stricter TAR@FAR= 10^{-4} protocol.

An intriguing observation in Table II is that while the three-phase strategy (w-3p) is crucial for stabilizing smaller architectures like iResNet-18, the high-capacity iResNet-50 model (43.6M parameters) achieves its peak performance of 56.233% without the three-phase schedule (wo-3p), exceeding the full-precision teacher's performance (53.533%).

This phenomenon stems from the interplay between model capacity and data richness. The deep representation space of iResNet-50 allows it to learn robust biometric features directly from the 2,000 pose-rich VGGFace2 images in our refined dataset. Since the teacher model itself exhibits limited performance on extreme viewpoints, the strict trajectory alignment in the w-3p setting acts as an over-constraint. Providing optimization "freedom" in the wo-3p configuration allows the high-capacity student to learn directly from rich real-world pose variations, enabling it to surpass the teacher's embedding manifold limitations.

C. Convergence Efficiency

The proposed framework demonstrates a substantial improvement in training efficiency compared to existing synthetic-data-based approaches. By leveraging a refined dataset of approximately 53,500 real-world images and coordinating the optimization trajectory through the β-MSE objective, the student model effectively stabilizes its performance within a compact budget of 30 epochs.

To ensure a fair comparison with the state-of-the-art QuantFace [5], which utilizes a total training budget of 180,000 iterations with a batch size of 512, we normalize our computational requirements to the same batch size. Under this unified metric, our framework completes its 30-epoch fine-tuning schedule in only 3,135 iterations (compared to 12,540 iterations at our native batch size of 128).

This represents a 57.4$\times$ faster convergence rate while achieving superior recovery performance, such as the 102% recovery rate on AgeDB-30. These results confirm that a curated real-world dataset, when combined with gradient coordination, can provide improved data efficiency for low-bit face recognition under our experimental setting compared with scaling synthetic data volume.

D. Relative Gap Analysis

To provide an objective assessment of performance recovery across diverse evaluation protocols, we analyze the relative TAR gap. This metric complements the recovery rate by reporting the absolute TAR difference in percentage points, ensuring a fair comparison between our framework and synthetic-data-based SOTA like *QuantFace* despite their use of different test sets (e.g., IJB-B/C vs. LFW/CFP/AgeDB). The gap is defined as:

$$\textit{Gap} = TAR_{\text{Teacher}} - TAR_{\text{Student,Q6}}. \tag{8}$$

Table IV summarizes these relative gaps. While smaller positive values indicate near-teacher performance, negative values represent cases where the Q6 student surpasses the teacher. Remarkably, our models achieve significant negative gaps, such as -2.933 on AgeDB-30 and -2.700 on CPLFW.

TABLE IV
RELATIVE GAP (TAR@FAR=10^{-4}) COMPARISON ACROSS BENCHMARKS.

Architecture	Relative Gap (TAR@FAR=10^{-4})						
	IJB-C	IJB-B	LFW	CFP-FP	AgeDB-30	CALFW	CPLFW
iResNet-18 (real) [5]	0.53	0.56	–	–	–	–	–
iResNet-50 (real) [5]	4	4.12	–	–	–	–	–
mobilefacenet (real) [5]	7.75	8.01	–	–	–	–	–
iResNet-18 (w-3p)	–	–	-0.100	1.143	-2.933	-2.267	-2.367
iResNet-50 (wo-3p)	–	–	0.167	7.400	4.600	0.067	-2.700
mobilefacenet (w-3p)	–	–	2.267	2.943	30.100	11.500	5.567

These results suggest that the student model does not merely replicate the teacher and may benefit from complementary representations learned from the refined real-world pose and age subsets. Furthermore, achieving these competitive gaps with 57.4× faster convergence (3,135 vs. 180,000 iterations when normalized to the same batch size) demonstrates that our gradient coordination strategy provides a more stable and efficient optimization path than large-scale real data scaling [5].

VI. Conclusion

This paper presents a practical framework for low-bit face recognition based on Q6 quantization, teacher-guided embedding alignment, and refined real-data training. The experimental results show that the proposed student model remains lightweight enough for resource-constrained deployment while still preserving strong recognition capability across standard benchmarks. In particular, the Q6 model achieves performance close to the full-precision teacher and even exceeds it on AgeDB-30, demonstrating that an efficient low-bit model can still maintain highly discriminative face embeddings.

From an application perspective, these findings indicate that the proposed method is well suited for real-world scenarios where memory, computation, and latency are limited, such as embedded systems, mobile devices, and edge AI platforms. Rather than requiring massive synthetic augmentation or costly large-scale retraining, the framework enables a compact quantized model to recover substantial performance using a much smaller curated dataset. This makes the approach more practical for deployment settings in which both computational resources and training data budgets are constrained.

More importantly, this study suggests a methodological insight: for low-bit face recognition, improving the optimization strategy is a more effective direction than merely increasing data scale. By combining gradient coordination, staged quantization-aware training, and refined real-world data variations, the proposed method shows that algorithmic design can play an important role in closing the gap between quantized and full-precision models. Most notably, the proposed framework achieves 57.4× faster convergence, requiring only 3,135 iterations compared to the 180,000 iterations required by real-data-based state-of-the-art approaches. This suggests that future progress in efficient face recognition should continue to focus on stronger optimization objectives and better training dynamics, rather than relying only on larger datasets.

Appendix

Due to page limitations, details addressing the query on hyperparameter selection (specifically the sensitivity analysis and selection justification for the varying β values across different architectures and benchmarks), along with additional ablation studies, multi-seed results, and detailed implementations, are available at https://github.com/thien1234ff/6-bit_FaceRecognition.git.

Acknowledgment

This work was supported by the University Research Grant from the University of Sciences, Hue University, under Project No. DHKH2026B-002.

References

[1] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2019.

[2] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, and W. Liu, "CosFace: Large margin cosine loss for deep face recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2018.

[3] B. Jacob *et al.*, "Quantization and training of neural networks for efficient integer-arithmetic-only inference," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2018.

[4] S. Zhou, Y. Wu, Z. Ni, X. Zhou, H. Wen, and Y. Zou, "DoReFa-Net: Training low bitwidth convolutional neural networks with low bitwidth gradients," *arXiv preprint arXiv:1606.06160*, 2016.

[5] Boutros, F., Damer, N., & Kuijper, A., "QuantFace: Towards Lightweight Face Recognition by Synthetic Data Low-Bit Quantization," in *Proceedings of the 26th International Conference on Pattern Recognition (ICPR)*, pp. 855–862, IEEE, Aug. 2022.

[6] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2016.

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

[8] A. Romero *et al.*, "FitNets: Hints for thin deep nets," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2015.

[9] A. G. Howard *et al.*, "MobileNets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[10] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2018.

[11] C.-P. Tran, A.-K. N. Vu, and V.-T. Nguyen, "Baby learning with vision transformer for face recognition," in *Proc. Int. Conf. Multimedia Analysis and Pattern Recognition (MAPR)*, 2022.

[12] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," University of Massachusetts, Amherst, Tech. Rep. 07-49, Oct. 2007.

[13] S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia, and S. Zafeiriou, "AgeDB: The first manually collected, in-the-wild age database," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshops (CVPRW)*, Honolulu, HI, USA, Jul. 2017, pp. 1997–2005.

[14] S. Sengupta, J. Chen, C. D. Castillo, V. M. Patel, R. Chellappa, and D. W. Jacobs, "Frontal to profile face verification in the wild," in *Proc. IEEE Winter Conf. Applications of Computer Vision (WACV)*, Lake Placid, NY, USA, Mar. 2016, pp. 1–9.

[15] T. Zheng, W. Deng, and J. Hu, "Cross-age LFW: A database for studying cross-age face recognition in unconstrained environments," *CoRR*, vol. abs/1708.08197, 2017. [Online]. Available: http://arxiv.org/abs/1708.08197

[16] T. Zheng and W. Deng, "Cross-pose LFW: A database for studying cross-pose face recognition in unconstrained environments," Beijing University of Posts and Telecommunications, Tech. Rep. 18-01, Feb. 2018.

[17] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning Face Representation from Scratch," *arXiv preprint arXiv:1411.7923*, 2014.

[18] R. Rothe, R. Timofte, and L. Van Gool, "DEX: Deep EXpectation of Apparent Age from a Single Image," in *Proc. IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2015, pp. 10–15.

[19] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFace2: A Dataset for Recognising Faces across Pose and Age," in *Proc. IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 2018, pp. 67–74.