

Article

FD-GhostFaceNet: Frequency-Decoupled Ghost Modules for Lightweight Face Recognition

Abdalbasit Qadir ^{1,*}, Bryar A. Hassan ^{1,2}  and Hozan Khalid ^{3,4}

¹ Computer Science Department, College of Science, Charmo University, Chamchamal 46023, Iraq; bryar.ahmad@ukh.edu.krd

² Computer Science and Engineering Department, School of Science and Engineering, University of Kurdistan Hewler, Erbil 44001, Iraq

³ Information Technology Department, Computer Science Institute, Sulaimani Polytechnic University, Sulaymaniyah 46001, Iraq; hozan.khalid@spu.edu.iq

⁴ Computer Engineering Department, Tishk International University, Erbil 44001, Iraq

* Correspondence: abdulbast.mhamad@chu.edu.iq

Abstract

Deploying accurate face recognition models on resource-constrained devices remains a significant challenge. Ghost modules mitigate feature-map redundancy by synthesizing half of each feature map through inexpensive linear transformations and form the basis of the lightweight GhostFaceNet and GhostFaceNet++ families. Nevertheless, the inexpensive generator is a strictly local 3×3 depthwise convolution: the synthesized features remain near-duplicates of the intrinsic ones, and deeper stages lack the global receptive field required for cross-pose and cross-age matching. This paper introduces FD-GhostFaceNet, built upon the proposed Frequency-Decoupled Ghost (FD-Ghost) module, which replaces the local generator with a complementary-band global counterpart. A learnable, input-conditioned partition of the two-dimensional Discrete Cosine Transform spectrum decomposes the intrinsic features into a global low-frequency band that encodes pose- and age-stable identity structure and a locally refined high-frequency detail band. Nyquist-consistent anti-aliased downsampling, which band-limits each feature map before subsampling so that the decoupled high-frequency band survives decimation, completes the frequency-decoupled design. Applied to both published trunks, FD-GhostFaceNet-V1-2 and FD-GhostFaceNet-V2-2 require 4.60 M and 7.39 M parameters, 83.1 and 97.1 MFLOPs, and 9.19 and 14.77 MB of storage, respectively, and thus fall within the sub-100 MFLOPs category of lightweight face recognition models. Across six standard benchmarks, both variants surpass their corresponding GhostFaceNet baselines in 23 of 24 trunk-benchmark comparisons and are competitive with or superior to the GhostFaceNet++ variants, with the largest gains on the pose-sensitive evaluations. When trained with ArcFace on UMDFaces, FD-GhostFaceNet raises the best reported CP-LFW accuracy from 84.65% to 86.12% and CFP-FP from 87.60% to 89.56%; when trained on CASIA-WebFace, it improves the best reported CFP-FP from 90.10% to 90.94%. These results confirm that frequency decoupling strikes a favorable balance between compactness and accuracy, making both models strong candidates for deployment on resource-constrained devices.

Keywords: face recognition; lightweight deep learning; Ghost networks; biometric recognition; frequency-decoupled learning



Academic Editor: Mariofanna Milanova

Received: 17 June 2026

Revised: 30 July 2026

Accepted: 5 August 2026

Published: 7 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article

distributed under the terms and

conditions of the [Creative Commons](#)

[Attribution \(CC BY\)](#) license.

1. Introduction

Face recognition has achieved near-human or even super-human accuracy on standard benchmarks over the past decade, driven by deep convolutional neural networks (CNNs) trained with large-scale datasets and discriminative margin-based loss functions. Additive angular margin loss (ArcFace) [1] and large margin cosine loss (CosFace) [2] have established state-of-the-art performance on unconstrained verification tasks by maximizing inter-class separability on a hypersphere, yielding verification accuracies above 99.8% on the Labeled Faces in the Wild (LFW) benchmark [3]. Despite these advances, translating such accuracy to mobile, embedded, and edge-computing platforms remains a persistent challenge. Devices in these deployment contexts operate under tight budgets of memory, multiply–accumulate operations (MACs), and battery energy, while the large CNNs responsible for state-of-the-art performance routinely require hundreds of megabytes of storage and billions of FLOPs per inference. The lightweight face recognition line has sought to address this gap: MobileFaceNet [4] distills face-specific priors into a depthwise separable backbone at approximately 439.8 MFLOPs, while subsequent architectures, including ShuffleFaceNet [5] and VarGFaceNet [6], explore grouped and variable-group convolutions to trade accuracy for efficiency. Nevertheless, a meaningful accuracy gap between these compact models and their large-model counterparts persists, particularly under challenging conditions such as large pose variation and cross-age verification.

The Ghost module family offers a principled route to compactness by exploiting the well-documented redundancy in CNN feature maps. GhostNet [7] demonstrated that roughly half of each feature map can be synthesized from the other half through cheap linear transformations, specifically 3×3 depthwise convolutions, rather than via full convolutions, reducing both parameter count and FLOPs with negligible accuracy degradation. GhostNetV2 [8] extended this framework with a decoupled fully connected (DFC) attention mechanism that aggregates long-range spatial context through horizontal and vertical fully connected layers applied to a down sampled feature map. Adapting these architectures to face recognition, GhostFaceNets [9] produced a family of face verification models spanning 60–275 MFLOPs that achieved state-of-the-art results among lightweight models when trained with ArcFace. GhostFaceNet++ [10] further refined the design by incorporating Cross-Stage Partial (CSP) connections and Efficient Channel Attention (ECA), its strongest (0.5-ratio) variants spanning 52–62 MFLOPs while improving verification accuracy across multiple benchmarks.

Notwithstanding these advances, two structural limitations remain unresolved in the Ghost-based face recognition line. First, the inexpensive generator that synthesizes the second half of each Ghost feature map is a strictly local 3×3 depthwise convolution; because the filter footprint is small and the operation is channel-separable, the synthesized features remain near-duplicates of the intrinsic features, providing limited additional discriminative capacity beyond a narrow spatial neighborhood. Second, the synthesized path contains no operator with a global receptive field in the deep, low-resolution stages (from the 28×28 down to the 4×4 spatial grids) of the backbone. Global facial geometry, the relative configuration of eyes, nose, and jawline encoded across the full feature map, is essential for robust cross-pose matching, as evidenced by the persistent difficulty of lightweight models on CP-LFW [11] and CFP-FP [12], and for age-invariant matching on CA-LFW [13] and AgeDB-30 [14]. The strongest GhostFaceNet++ variants report 84.65% on CP-LFW and 87.60% on CFP-FP under the UMDFaces protocol, indicating that pose remains the weakest axis even after incorporating CSP structure and ECA. GhostNetV2's DFC attention partially addresses the locality problem, but its gate is computed on a stride-2 average-pooled map; at the 7×7 and 4×4 resolutions of the deep stages, the gate is therefore estimated from a 3×3 or 2×2 grid, far too coarse to represent facial

configuration, rendering the long-range mechanism ineffective precisely where global structural information is most needed.

Frequency-domain representations offer a theoretically grounded and computationally efficient means of accessing global context in feature maps. The two-dimensional Discrete Cosine Transform (DCT) concentrates signal energy in a small number of low-frequency coefficients [15], a property that underlies JPEG compression and has motivated growing adoption in deep learning. FcaNet [16] demonstrated that global average pooling is a special case of the DC (lowest-frequency) component of the 2D DCT and enriched channel attention with multiple DCT components, with the coefficients used solely to produce per-channel scalar attention weights. GFNet [17] replaced self-attention in vision transformers with learnable global filters applied via the 2D FFT, achieving log-linear token mixing through complex-valued spectral operations over the full frequency plane. Octave convolution [18] factorizes feature maps into a low-frequency component stored at half resolution and a high-frequency residual using a fixed, non-learnable partition. Each of these prior works exploits the frequency domain for a distinct purpose, namely channel attention scalars, full token mixing, or fixed multi-resolution factorization; none integrate a learnable, input-conditioned frequency-band partition directly into the feature generation step of a Ghost module, nor condition the partition on the input.

The present work introduces FD-GhostFaceNet, built upon the proposed Frequency-Decoupled Ghost (FD-Ghost) module, which replaces the local depthwise generator of the standard Ghost module with a global, complementary-band counterpart. The intrinsic features are first projected to the frequency domain via an orthonormal 2D DCT. A learnable per-channel, per-frequency gate, augmented with an input-conditioned per-channel shift of the band boundary, then partitions the spectrum into a global low-frequency band, encoding pose- and age-stable identity structures by virtue of its full-map receptive field and a complementary high-frequency band, subsequently refined by a lightweight local depthwise convolution to recover fine-grained texture detail. The two bands sum exactly to the intrinsic features, guaranteeing perfect reconstruction and hence no representational loss at initialization. To protect the most stable identity information, the DC coefficient is explicitly anchored to the low-frequency band regardless of the learned gate value, and a bounded cross-band bilinear interaction between the two bands introduces second-order representational capacity without increasing architectural depth. Complementing the within-stage partition, anti-aliased BlurPool downsampling [19] is applied at the deep stride-2 stages: band-limiting each feature map before subsampling ensures that the high-frequency band decoupled by FD-Ghost survives decimation, improving shift invariance under small geometric perturbations. The resulting models operate at 4.60 M parameters, 83.1 MFLOPs, and 9.19 MB of fp16 storage (FD-GhostFaceNet-V1-2), and 7.39 M parameters, 97.1 MFLOPs, and 14.77 MB (FD-GhostFaceNet-V2-2), both well below MobileFaceNet's 439.8 MFLOPs, with the V1 variant requiring fewer parameters and a smaller storage footprint than the GhostFaceNet V2-2 baseline (6.84 M parameters, 13.68 MB).

Extensive experiments on six standard face verification benchmarks under two training protocols confirm the effectiveness of the proposed approach. When trained with ArcFace on UMDFaces [20], FD-GhostFaceNet surpasses its corresponding GhostFaceNet baseline on all six benchmarks and sets new best results among all GhostFaceNet and GhostFaceNet++ variants on the pose-sensitive benchmarks, raising CP-LFW from 84.65% to 86.12% and CFP-FP from 87.60% to 89.56% while remaining within statistical parity of the strongest variants on the saturated frontal benchmarks. When trained on CASIA-WebFace [21], the model again surpasses its corresponding baseline on eleven of the twelve comparisons (the single exception, AgeDB-30 for V2-2, trails by only 0.07 points), most notably improving CFP-FP from 90.10% to 90.94% and raising the best reported LFW

accuracy from 99.05% to 99.08%, with CP-LFW and CFP-FF within statistical parity of the strongest variants. The principal contributions of this work are as follows:

- **Frequency-Decoupled Ghost (FD-Ghost) module:** A novel Ghost module that replaces the strictly local 3×3 depthwise generator with an orthonormal 2D DCT spectral decomposition controlled by a learnable, input-conditioned per-channel gate, partitioning the spectrum into a global low-frequency band and a complementary high-frequency band—summing exactly to the intrinsic features—with the high band subsequently refined by a lightweight local depthwise convolution, thereby endowing the synthesized features with a full-map receptive field at the deep stages where existing long-range mechanisms degenerate.
- **DC anchoring and cross-band bilinear interaction:** An explicit constraint that assigns the DC (global-mean) coefficient unconditionally to the stable low-frequency band to preserve identity structure, combined with a bounded bilinear interaction between the two complementary bands that introduces second-order representational capacity without increasing architectural depth.
- **Nyquist-consistent anti-aliased downsampling:** BlurPool downsampling at the deep stride-2 stages band-limits each feature map before subsampling, so that the high-frequency band decoupled by the FD-Ghost partition survives decimation between stages; this completes the frequency-decoupled design and improves shift invariance under the small geometric perturbations characteristic of profile-pose alignment.
- **State-of-the-art pose accuracy among Ghost-based lightweight face recognition models:** At 4.60 M parameters/83.1 MFLOPs (V1) and 7.39 M parameters/97.1 MFLOPs (V2), FD-GhostFaceNet surpasses its corresponding GhostFaceNet baseline in 23 of 24 trunk-benchmark comparisons across both training protocols (the single exception is within 0.1 point) and sets new best results among all GhostFaceNet and GhostFaceNet++ variants on CP-LFW (86.12%) and CFP-FF (89.56%) under UMDFaces and on CFP-FF (90.94%) and LFW (99.08%) under CASIA-WebFace, demonstrating that frequency decoupling within the Ghost generation step yields a favorable balance between compactness and accuracy for resource-constrained deployment.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 describes the FD-Ghost module and the full FD-GhostFaceNet architecture in detail. Section 4 presents the experimental setup, results and benchmark comparisons. Section 5 concludes the paper and outlines directions for future research.

2. Related Works

2.1. Lightweight and Ghost-Based Face Recognition

The design of compact convolutional neural networks (CNNs) for resource-constrained hardware has been an active research area since the mid-2010s. SqueezeNet [22] pioneered extreme compactness through Fire modules, which squeeze channels with 1×1 convolutions before a mixed $1 \times 1/3 \times 3$ expansion, attaining AlexNet-level ImageNet accuracy with a model size below 0.5 MB after compression. MobileNetV1 [23] systematically replaced standard convolutions with depthwise separable convolutions, factoring each spatial filter from its pointwise projection, and MobileNetV2 [24] introduced inverted residual blocks with linear bottlenecks, expanding intermediate channels for non-linear filtering while preserving capacity in the narrow layers. ShuffleNetV1 [25] and ShuffleNetV2 [26] demonstrated that group convolutions combined with channel shuffling can match the efficiency of depthwise designs while achieving superior practical throughput through hardware-aligned memory access patterns.

Adapting these general-purpose backbones to face recognition requires task-specific design choices. MobileFaceNet [4] replaced the global average pooling head with a Global

Depthwise Convolution head, providing spatially weighted feature aggregation tailored to the 112×112 aligned face crop; the resulting model achieves competitive accuracy at approximately 439.8 MFLOPs and remains a widely used efficiency reference. ShuffleFaceNet [5] adapted the ShuffleNetV2 block to face verification, VarGFaceNet [6] introduced variable-group convolutions to balance channel width and filter granularity, MixFaceNets [27] employed mixed-kernel depthwise convolutions to capture multiple spatial scales within a single block, and PocketNet [28] combined neural architecture search with multi-step knowledge distillation to obtain face-specific topologies with as few as 0.92 million parameters. More recently, EdgeFace [29] combined CNN and transformer components with low-rank linear layers, ranking first in the IJCB 2023 Efficient Face Recognition Competition. Despite these advances, models in this family exhibit disproportionate accuracy degradation on cross-pose benchmarks, such as CP-LFW and CFP-FP, and on cross-age benchmarks, such as CA-LFW and AgeDB-30, relative to their performance on the near-frontal LFW set, a gap attributable in part to the limited receptive fields of the efficient operators they employ.

The GhostNet family [7] departs from these channel-reduction strategies by explicitly exploiting the redundancy present in CNN feature maps. Rather than computing all output channels through full convolutions, a Ghost module first produces a smaller set of intrinsic feature maps via a conventional 1×1 convolution, then synthesizes a complementary set of Ghost maps by applying inexpensive per-channel linear transformations, implemented as 3×3 depthwise convolutions, to the intrinsic maps; concatenating the two sets recovers the original channel count at reduced compute and parameter cost. This plug-and-play bottleneck achieves 75.7% top-1 accuracy on ImageNet, exceeding MobileNetV3 at a comparable computational budget. GhostNetV2 [8] identified a residual weakness of this design: the 3×3 depthwise generator is spatially local and cannot encode the long-range dependencies required for robust recognition. It therefore augments each Ghost bottleneck with a decoupled fully connected (DFC) attention gate, which applies stride-2 average pooling to the input, computes decoupled horizontal and vertical fully connected projections over the resulting low-resolution map, and produces a multiplicative gate that is upsampled back to the original resolution before element-wise scaling of the Ghost output. GhostNetV3 [30] retains this topology but improves accuracy through structural re-parameterization and multi-step knowledge distillation, without altering the inference-time architecture.

Building on these backbones, GhostFaceNets [9] transferred GhostNetV1 and GhostNetV2 to face verification by pairing them with a Global Depthwise Convolution (GDC) embedding head and training with the ArcFace loss, achieving state-of-the-art accuracy among lightweight models at 60–275 MFLOPs, substantially below MobileFaceNet. GhostFaceNet++ [10] further refined the family by applying Cross-Stage Partial (CSP) partitioning within the Ghost bottleneck to reduce redundant computation and by inserting an Efficient Channel Attention (ECA) module in the embedding head to recalibrate channel importance at negligible parameter overhead; these modifications compressed the operating range to 51–62 MFLOPs and raised the best reported CP-LFW accuracy to 84.65% and CFP-FP to 87.60% under the UMDFaces training protocol.

Throughout this line, the training objective has converged on margin-based angular losses. Within the lightweight face recognition literature, ArcFace with a feature scale of 64 and an angular margin of 0.5 has become the de facto standard: GhostFaceNets, GhostFaceNet++, and the present work all adopt this identical configuration, ensuring that accuracy differences are attributable solely to the backbone architecture. The proposed contribution is therefore orthogonal to the choice of margin loss.

Despite these advances, a structural limitation persists across all existing Ghost variants: the cheap branch that generates the synthesized half of the feature maps remains an

exclusively local operator. From GhostNetV1 through GhostFaceNet++, this generator is a 3×3 depthwise convolution whose receptive field is confined to a nine-position neighborhood. The DFC attention of GhostNetV2 does provide long-range context, but it acts as a multiplicative gate applied to the generated features rather than through the generation process itself; moreover, at the deep 7×7 and 4×4 stages that encode full-face geometry, its stride-2 downsampling reduces the gate's context map to a 3×3 or 2×2 grid, severely constraining the spatial structure it can represent. In contrast, the proposed FD-Ghost module makes the generated half itself globally informed: an orthonormal 2D DCT spectral partition with full-map receptive support replaces the local generator at the deep stages (28×28 and coarser) where global facial structure is encoded and where the existing long-range mechanism is weakest.

2.2. Frequency-Domain Methods in Computer Vision

The Discrete Cosine Transform (DCT) is an orthogonal linear transform that concentrates the energy of natural signals into a small number of low-frequency coefficients, a property known as energy compaction [15] that underpins the JPEG compression standard. In deep learning, the frequency domain has been exploited along several distinct lines.

Channel attention from DCT coefficients [16] showed that global average pooling, the spatial compression step of squeeze-and-excitation attention, is equivalent to retaining only the DC component of the 2D DCT, namely the per-channel spatial mean. FcaNet generalizes this observation by selecting multiple DCT components as richer channel descriptors, which are then passed through the standard gating to produce per-channel scalar attention weights. The DCT is used solely to generate compact scalar descriptors; the spatial feature maps themselves are neither partitioned nor modified by frequency.

Spectral token mixing: GFNet [17] replaced self-attention in vision transformers with learnable complex-valued filters applied through the 2D FFT, element-wise multiplication in the frequency domain, and an inverse FFT, achieving log-linear token mixing. The Adaptive Fourier Neural Operator [31] extended this principle with block-diagonal channel mixing and soft-thresholding of frequency modes, again mixing across the entire frequency plane through complex Fourier coefficients.

Fixed multi-resolution factorization: Octave convolution [18] factorizes feature maps into a low-frequency component stored at half spatial resolution and a high-frequency residual at full resolution, with defined inter-scale information flow. The low/high partition is determined by a fixed downsampling criterion; it is neither learnable nor input-adaptive.

Frequency-domain learning on compressed inputs: Networks operating directly on DCT coefficients decoded from JPEG can match spatial-domain accuracy while avoiding full decompression [32], and wavelet-integrated CNNs [33] apply multi-scale wavelet decompositions to improve noise robustness. These approaches modify the network's input representation rather than any internal feature-generation mechanism.

Anti-aliasing as a frequency-domain concern: BlurPool [19] demonstrated that strided downsampling violates the Nyquist criterion, causing shift sensitivity in deep networks, and that inserting a fixed low-pass filter before each subsampling step restores shift invariance. The present work adopts this rationale directly: BlurPool is applied at the deep downsampling sites of FD-GhostFaceNet, where it complements the FD-Ghost module by band-limiting feature maps between stages while FD-Ghost partitions frequency content within each stage.

Frequency-aware face analysis: In face forgery detection, spectral artifacts introduced by generative models are exploited in the DCT or DFT domain [34,35]. Such methods detect frequency anomalies in input images to classify authenticity, a fundamentally different

objective from face verification, where both probes are genuine images and the task is identity matching across pose and age variation.

Taken together, prior frequency-domain methods use the spectrum as a source of channel attention scalars (FcaNet), a medium for full-plane token mixing (GFNet, AFNO), a fixed two-scale spatial factorization (octave convolution), a compressed input representation (DCT- and wavelet-domain networks), or a band-limiting filter for downsampling (BlurPool). None embed a learnable, input-conditioned frequency-band partition into the feature-generation step of an efficiency-oriented convolution module, and none couple that partition to the distinction between identity-stable low-band structure and identity-variable high-band detail under an exact-reconstruction guarantee. This is precisely the gap that the proposed FD-Ghost module fills.

3. Proposed Approach

This section details FD-GhostFaceNet. We first review the Ghost module and the GhostFaceNet baseline conventions that our backbone inherits (Section 3.1), then introduce the Frequency-Decoupled Ghost module (Section 3.2), the bottleneck and overall architecture built upon it (Section 3.3), and the Nyquist-consistent anti-aliased downsampling that completes the frequency-decoupled design (Section 3.4).

3.1. Preliminaries: Ghost Modules and the GhostFaceNet Baseline

A Ghost module [7] reduces the cost of feature generation by computing only half of its output channels with a full operator and synthesizing the remainder cheaply. Given an input tensor x , an intrinsic half P is produced by a pointwise convolution followed by batch normalization and PReLU, and a Ghost half is synthesized from P by a 3×3 depthwise convolution with the same normalization and activation; the module output is the channel-wise concatenation of the two halves. GhostNetV2 supplements this design with a decoupled fully connected (DFC) attention gate computed on a downsampled feature map. GhostFaceNets [9] adapt these backbones to face recognition by replacing ReLU with PReLU, implementing the squeeze-and-excitation (SE) modules with convolutions at reduction ratio 0.25, reading the embedding through a Global Depthwise Convolution (GDC) head, and training with the ArcFace loss.

The backbones considered in this work inherit these conventions; both published trunks are used, the V1-2 backbone and the V2-2 backbone, whose DFC attention is retained verbatim. The input is a 112×112 aligned face crop; a stride-2 stem convolution yields a 56×56 feature map, and the subsequent stages reduce resolution as $56 \rightarrow 28 \rightarrow 14 \rightarrow 7 \rightarrow 4$. A width multiplier $\alpha = 1.3$ is applied throughout, the final pointwise convolution produces 664 channels on the V1 trunk and 1248 on the V2 trunk, and the embedding is 512-dimensional. Two structural limitations of this baseline motivate our design. First, the cheap generator is strictly local: its 3×3 depthwise receptive field cannot synthesize global structure, so the Ghost half remains a near-duplicate of the intrinsic half. Second, no operator in the generation path attains a global receptive field at the deep 28×28 , 14×14 , 7×7 , and 4×4 stages, where full-face geometry is encoded and where the DFC gate degenerates: at the 7×7 and 4×4 stages its stride-2 pooled context map shrinks to only 3×3 or 2×2 cells.

3.2. Frequency-Decoupled Ghost (FD-Ghost) Module

The FD-Ghost module in Figure 1 replaces the local cheap generator with a complementary-band global one. The intrinsic half P is computed exactly as in the standard Ghost module by a pointwise convolution followed by batch normalization and PReLU

(the activation is omitted in the projection Ghost module of each bottleneck, following the baseline).

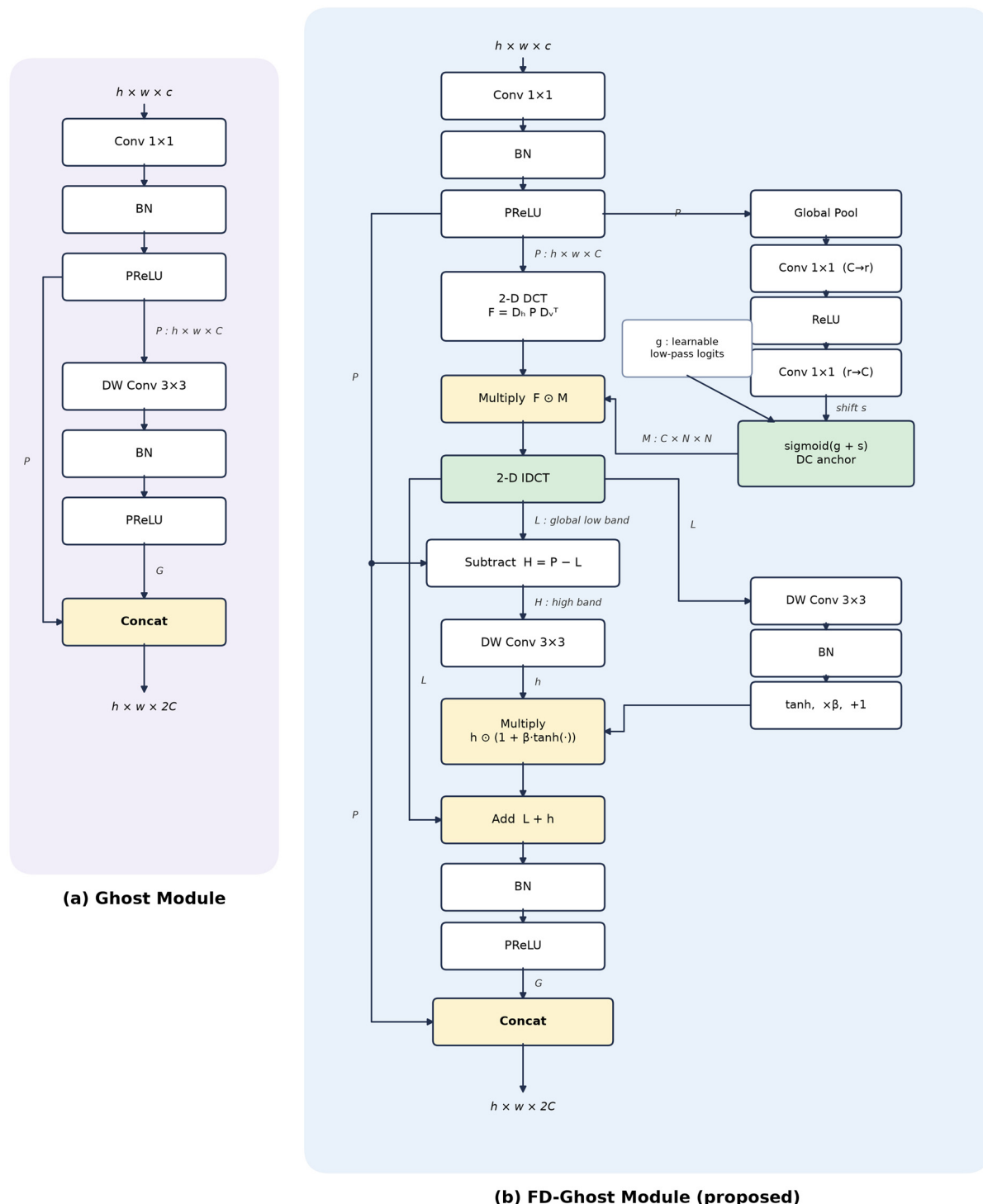


Figure 1. Ghost module vs. FD-Ghost module.

Each stage of spatial size $N \in \{28, 14, 7, 4\}$ is equipped with a fixed orthonormal DCT-II basis $D \in \mathbb{R}^{N \times N}$ satisfying $DD^T = I$, so the two-dimensional spectrum of each channel of P is obtained as $F = DhP Dv^T$, where $Dh = Dv = D$ denote the row- and column-wise application of D (the feature maps are square), the inverse transform is simply the transpose,

and the transform is energy-preserving with no normalization overhead. A soft low-band selector is computed per channel c and frequency (u, v) as

$$M_{c,u,v} = \text{sigmoid}(g_{c,u,v} + s_c(P)), \quad (1)$$

where g are learnable logits initialized to a smooth radial low-pass profile, $g^0(u,v) = \text{clip}(2 - 4 \cdot \sqrt{((u/(N-1))^2 + (v/(N-1))^2)}, -4, 4)$, which is positive at DC and negative toward the spectral corner, and $s_c(P)$ is an input-conditioned per-channel shift produced by a lightweight squeeze module (global average pooling, a pointwise convolution to a reduced width of at most 16 channels—fixed a priori and not tuned—a ReLU, and a second pointwise convolution). The shift moves the band boundary per channel and input, allowing the band boundary to vary with the input; it does not generate input-dependent convolution kernels, which distinguishes the mechanism from dynamic convolution. To protect the most pose- and age-stable component, the DC coefficient, which is proportional to the per-channel spatial mean, is unconditionally anchored to the low band by overriding the gate at the $(0, 0)$ frequency, an operation with no computational cost. The gated spectrum is mapped back to the spatial domain to form the global low-frequency band, and the complementary high-frequency band follows by exact subtraction:

$$L = D_h^\top (F \odot M) D_v, \quad H = P - L. \quad (2)$$

Two properties are worth emphasizing. First, every DCT basis function has full spatial support, so L carries a full-map receptive field by construction, supplying precisely the global operator the Ghost generation path lacks. Second, $L + H = P$ holds for any gate value (perfect reconstruction, which also saves a second inverse transform), so the decomposition incurs no representational loss and the module is well conditioned at initialization.

The high band then receives local refinement by a 3×3 depthwise convolution, $h = \text{DWConv}_{3 \times 3}(H)$, modulated by a bounded cross-band bilinear interaction in which the stable low band selects the identity-relevant high-frequency detail: h is scaled element-wise by $1 + \beta_c \tanh(\text{BN}(\text{DWConv}_{3 \times 3}(L)))$, where β_c is a per-channel parameter initialized to zero, so the module starts as the purely additive form with no initialization regression, and the $\tanh$ bounds the modulation. The Ghost half is finally $G = \sigma(\text{BN}(L + h))$ with σ denoting PReLU, and the module output is the concatenation of P and G , exactly matching the output shape of the standard Ghost module.

The two halves thus carry complementary receptive fields and frequency content: P performs pointwise channel mixing, while G integrates global low-frequency identity structure with locally refined high-frequency detail. We make no claim that the halves are statistically decorrelated. Beyond the standard Ghost module, the added computation consists of the fixed DCT and inverse-DCT matrix multiplications underlying Equation (2), which amount to $2C(H^2W + HW^2)$ multiply-accumulate operations per module, where C denotes the channel count of P and the spatial size never exceeds 28×28 , together with the lightweight squeeze module and the depthwise convolution of the cross-band interaction. The added parameters comprise the gate logits, the squeeze module, and the cross-band depthwise convolution together with its batch normalization and per-channel gains β_c . All of these contributions are included in the reported parameter and FLOP totals.

3.3. FD-Ghost Bottleneck and Overall Architecture

The FD-Ghost bottleneck (Figure 2) retains the GhostNet bottleneck topology: an expansion Ghost module with PReLU, an optional stride-2 depthwise convolution (kernel 3 or 5 per the standard schedule, with BN and PReLU), an optional SE module (ratio 0.25), and a projection Ghost module without activation, closed by a residual connection. The

shortcut is the identity when input and output shapes match and a depthwise convolution with matching stride, pointwise convolution, and BN otherwise. On the V2 trunk, the DFC attention of GhostNetV2 is retained verbatim—attached to the expansion Ghost module from the third bottleneck onward, exactly as in the published baseline—with its sigmoid gate multiplied onto the FD-Ghost output wherever FD is active; FD-Ghost is therefore a pure addition to the published architecture rather than a replacement of any of its components. The V1 trunk contains no DFC, matching its published counterpart.

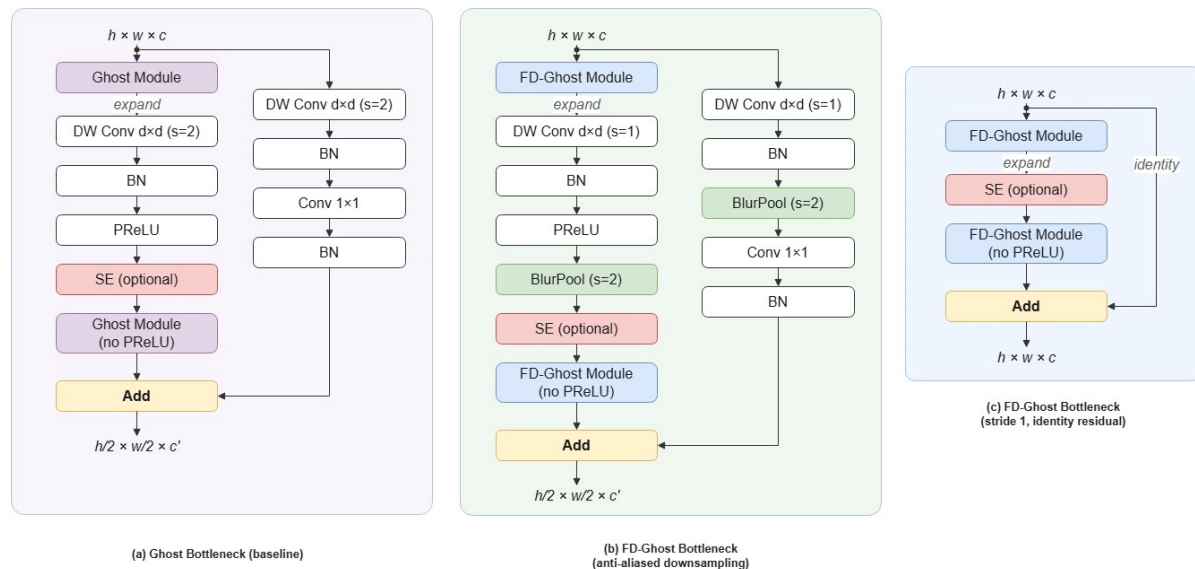


Figure 2. GhostNet bottleneck topology.

FD-Ghost generators are deployed where they matter and remain affordable: they replace the standard Ghost generators at all stages of spatial size 28 or smaller, while the 56×56 stage retains standard Ghost modules. This placement reflects that profile-pose facial geometry is encoded from the 28×28 stage downward, that DCT cost remains bounded because the transform matrices never exceed 28×28 , and that the DFC degeneration the module remedies occurs precisely at the deepest grids.

The overall network (Figure 3) comprises a 3×3 stride-2 stem convolution producing 20 channels, 16 bottlenecks distributed over the five stages following the standard GhostNet kernel, channel, and SE schedule at $\alpha = 1.3$, and a final pointwise convolution producing 664 channels (V1) or 1248 channels (V2) ahead of the standard GDC embedding head, which is unchanged from the baseline.

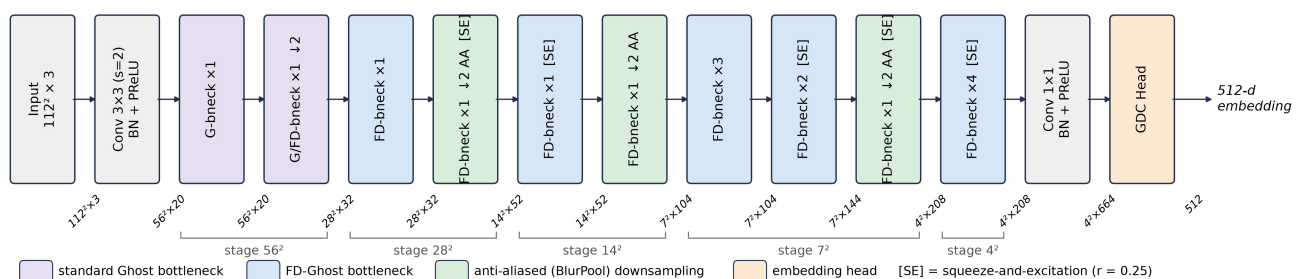


Figure 3. Overall network architecture.

3.4. Nyquist-Consistent Anti-Aliased Downsampling

Strided downsampling subsamples feature maps without band-limiting them, violating the Nyquist criterion and rendering deep representations sensitive to sub-pixel

input shifts—precisely the perturbations introduced by imperfect alignment of profile faces. Complementing the within-stage partition of Section 3.2, downsampling at the deep transitions is therefore anti-aliased: each stride-2 depthwise convolution, in both the main path and the downsampling shortcut, is replaced by its stride-1 counterpart followed by BlurPool [19], a fixed 3×3 binomial low-pass filter (the normalized outer product of [1,2]) applied depthwise with reflect padding and stride 2. Band-limiting both residual branches identically prevents the shortcut from re-injecting aliased content, and the operation adds no learnable parameters. On the V1 trunk, anti-aliasing is applied at the $28 \rightarrow 14$, $14 \rightarrow 7$, and $7 \rightarrow 4$ transitions; on the V2 trunk, whose DFC gate already applies average pooling at the shallower stages, it is restricted to the $14 \rightarrow 7$ and $7 \rightarrow 4$ transitions to avoid redundant low-pass smoothing. The two frequency mechanisms are complementary: FD-Ghost partitions frequency content within a stage, while BlurPool band-limits between stages so that the high-frequency band the partition decouples survives decimation.

FD-GhostFaceNet-V1-2 contains 4.60 M parameters (4.56 M trainable; totals include batch-normalization statistics, following the parameter-counting convention of [9,10]), requires 83.1 MFLOPs including all DCT and BlurPool computations (counted as multiply-accumulate operations, the FLOPs convention of [9,10]), and occupies 9.19 MB at fp16 precision; FD-GhostFaceNet-V2-2 contains 7.39 M parameters (7.29 M trainable), requires 97.1 MFLOPs, and occupies 14.77 MB. For comparison, the GhostFaceNet V1-2 baseline occupies 4.07 M parameters and 8.13 MB, the V2-2 baseline 6.84 M parameters and 13.68 MB, and ECA-CSP GhostFaceNet V2-2 5.83 M parameters and 11.66 MB.

4. Experiments and Results

4.1. Datasets

The proposed model is trained on the CASIA-WebFace [21] and UMDFaces [20] datasets, following the protocol of GhostFaceNet++ so that all comparisons are conducted under identical training data. CASIA-WebFace contains approximately 0.49 M Internet-collected images of 10,575 identities gathered under highly unconstrained indoor and outdoor conditions, with substantial variability in pose, expression, and illumination, and is a standard training corpus for deep face recognition. UMDFaces comprises 0.37 M web-collected images of 8277 identities and is characterized by particularly extensive pose variation, natural illumination changes, and frequent occlusion. All training images are aligned and cropped to 112×112 pixels following the standard InsightFace preprocessing.

Evaluation is performed on six complementary verification benchmarks that isolate the principal challenges of unconstrained face recognition. LFW provides general in-the-wild verification of moderate difficulty. CP-LFW and CFP-FP target cross-pose verification, the latter matching frontal against profile views, while CFP-FF provides the frontal–frontal counterpart on the same subjects. CA-LFW and AgeDB-30 target cross-age verification with age gaps of up to 30 years. Together, these benchmarks probe precisely the pose and age axes that motivate the frequency-decoupled design. Table 1 summarizes the training and evaluation datasets.

Table 1. Face recognition datasets used for training and evaluation.

Role	Dataset	Subjects	Images/Pairs	Verification Challenge	Reference
Training	CASIA-WebFace	10,575	494,414 images	Unconstrained, in-the-wild appearance variation	[21]
	UMDFaces	8277	367,888 images	Wide pose range under unconstrained capture	[20]

Table 1. Cont.

Role	Dataset	Subjects	Images/Pairs	Verification Challenge	Reference
Evaluation	LFW	5749	6000 pairs	General unconstrained face verification	[3]
	CP-LFW	3968	6000 pairs	Cross-pose with large yaw variation	[11]
	CA-LFW	4025	6000 pairs	Cross-age with intra-subject aging variation	[13]
	AgeDB-30	568	6000 pairs	Cross-age with ~30-year age gap	[14]
	CFP-FP	500	7000 pairs	Frontal-to-profile extreme pose change	[12]
	CFP-FF	500	7000 pairs	Frontal-to-frontal mild pose variation	[12]

4.2. Implementation Details

All models are implemented in PyTorch 2.8.0 with Python 3.12 and trained on a single NVIDIA GeForce RTX 5090 (NVIDIA Corporation, Santa Clara, CA, USA). All four training runs share one unified recipe, as described below.

Training uses the ArcFace loss with angular margin $m = 0.5$ and feature scale $s = 64$, with the scale warmed up at $s = 32$ during the first epoch. Optimization is performed with stochastic gradient descent (momentum 0.9) at batch size 256 for 50 epochs, under a cosine learning-rate schedule decaying from 0.1 to 10^{-5} . L2 regularization of 5×10^{-4} is applied with the magnitude matched to the Keras implementation of the baselines; BatchNorm parameters, biases, PReLU slopes, the FD-Ghost gate logits g and cross-band scales β , and the band-shift squeeze module are excluded from weight decay. Data augmentation is limited to random horizontal flipping, identical to the baselines. BatchNorm uses momentum 0.1 and $\epsilon = 10^{-5}$. For evaluation, embeddings of the original and horizontally flipped images are summed and L2-normalized, following the evaluation convention of the baselines' framework. The unified hyperparameters are listed in Table 2.

Table 2. Training hyperparameters (shared across both protocols).

Hyperparameter	Value
Input shape/embedding	$112 \times 112/512\text{-d}$
Loss	ArcFace, $m = 0.5$, $s = 64$ (warm-up $s = 32$, epoch 1)
Optimizer	SGD, momentum 0.9
L2 regularization	5×10^{-4}
Learning rate	$0.1 \rightarrow 10^{-5}$, cosine, 50 epochs
Batch size	256
Augmentation	Random horizontal flip
BatchNorm	momentum 0.1, $\epsilon = 10^{-5}$

4.3. Results

4.3.1. Performance on CASIA-WebFace

Table 3 reports the results of both FD-GhostFaceNet variants trained on CASIA-WebFace alongside the GhostFaceNet baselines and all GhostFaceNet++ variants; all models are trained on the same dataset with the same ArcFace configuration, so accuracy differences are attributable to the backbone architecture.

Table 3. Evaluation of the proposed FD-GhostFaceNet with all models are trained on the CASIA-WebFace dataset.

Model	FLOPs	Params	Size (MB)	LFW (%)	CP-LFW (%)	CA-LFW (%)	AgeDB-30 (%)	CFP-FP (%)	CFP-FF (%)
GhostFaceNet V1-2 (Baseline)	60,295,590	4,066,388	8.13	98.75	82.88	91.45	89.80	88.41	98.47
CSP GhostFaceNet V1-2 (0.25)	46,848,532	3,622,649	7.25	98.67	82.27	91.27	89.68	87.84	98.59
CSP GhostFaceNet V1-2 (0.5)	51,637,644	3,754,993	7.51	98.87	83.12	91.18	90.62	88.56	98.61
CSP GhostFaceNet V1-2 (0.75)	56,426,756	3,887,337	7.77	98.80	82.53	91.33	90.02	88.39	98.34
ECA-CSP GhostFaceNet V1-2 (0.5)	51,650,260	3,754,996	7.51	98.87	83.53	91.63	90.67	88.87	98.67
FD-GhostFaceNet-V1-2 (Ours)	83,097,592	4,595,322	9.19	98.82	84.03	92.13	90.38	90.26	98.57
GhostFaceNet V2-2 (Baseline)	76,513,640	6,839,104	13.68	99.05	84.13	91.28	90.75	89.29	98.57
CSP GhostFaceNet V2-2 (0.5)	61,976,534	5,828,405	11.66	99.03	84.38	91.92	91.07	89.76	98.83
ECA-CSP GhostFaceNet V2-2 (0.5)	62,000,246	5,828,408	11.66	98.88	84.53	91.82	91.05	90.10	98.76
FD-GhostFaceNet-V2-2 (Ours)	97,081,496	7,386,434	14.77	99.08	84.48	91.58	90.68	90.94	98.73

All models are trained with the identical ArcFace configuration ($m = 0.5$, $s = 64$); bold rows denote the proposed models.

Each FD-GhostFaceNet variant surpasses its corresponding GhostFaceNet baseline: FD-GhostFaceNet-V1-2 exceeds the V1-2 baseline on all six benchmarks, and FD-GhostFaceNet-V2-2 exceeds the V2-2 baseline on five, with AgeDB-30 statistically unchanged (-0.07). The largest improvement occurs on CFP-FP, which rises from 90.10% (ECA-CSP GhostFaceNet V2-2) to 90.94%, a gain of 0.84 percentage points on the benchmark that most directly probes the frontal-to-profile matching the global low-frequency pathway was designed to support; notably, both variants exceed every previously reported CFP-FP value, with the V1 variant reaching 90.26%—1.39 points above the strongest V1-family variant—at 4.60 M parameters. CP-LFW shows the same pattern within the V1 family, improving from 83.53% (ECA-CSP V1-2) to 84.03% (+0.50), while the V2 variant (84.48%) is statistically indistinguishable from the best reported 84.53%. On the age axis, the V1 variant raises CA-LFW from the previous overall best of 91.92% to 92.13% and improves AgeDB-30 by 0.58 points over its baseline (90.38% vs. 89.80%), while AgeDB-30 remains within 0.4 points of the strongest CSP variants. LFW rises marginally from the best reported 99.05% to 99.08%, and CFP-FF (98.73%) is within one standard error of the best reported 98.83%, indicating that the gains on the difficult pose benchmarks are obtained without sacrificing performance on the near-saturated frontal sets.

Figure 4 presents the ROC curves of FD-GhostFaceNet-V1-2 (a) and FD-GhostFaceNet-V2-2 (b) on the six CASIA-WebFace benchmarks, plotted on a logarithmic FAR axis to emphasize the low-FAR regime relevant to deployment. Both models exhibit the same three difficulty tiers. The near-frontal benchmarks are saturated, with LFW and CFP-FF reaching AUCs of 0.997–0.998 (EER 1.1–1.6%). The cross-age benchmarks and the frontal-to-profile benchmark form an intermediate tier (CFP-FP, AgeDB-30, CA-LFW; AUC 0.960–0.967, EER 8.7–11.0%), while the extreme cross-pose benchmark CP-LFW is hardest by a wide margin (AUC 0.899 for V1-2 and 0.906 for V2-2, EER 18.0% and 17.7%), its true acceptance rate collapsing at FAR = 10^{-3} . Comparing the two panels, the deeper V2-2 trunk shifts

precisely the two cross-pose curves upward—CFP-FP (AUC 0.960 $\rightarrow$ 0.963, EER 11.0% $\rightarrow$ 10.0%) and CP-LFW (AUC 0.899 $\rightarrow$ 0.906)—while the frontal tier is unchanged and CA-LFW marginally favors V1-2. This ordering—frontal verification essentially solved, cross-pose and cross-age remaining the binding constraints—is exactly the regime the global low-frequency pathway targets and identifies CP-LFW as the benchmark with the most headroom.

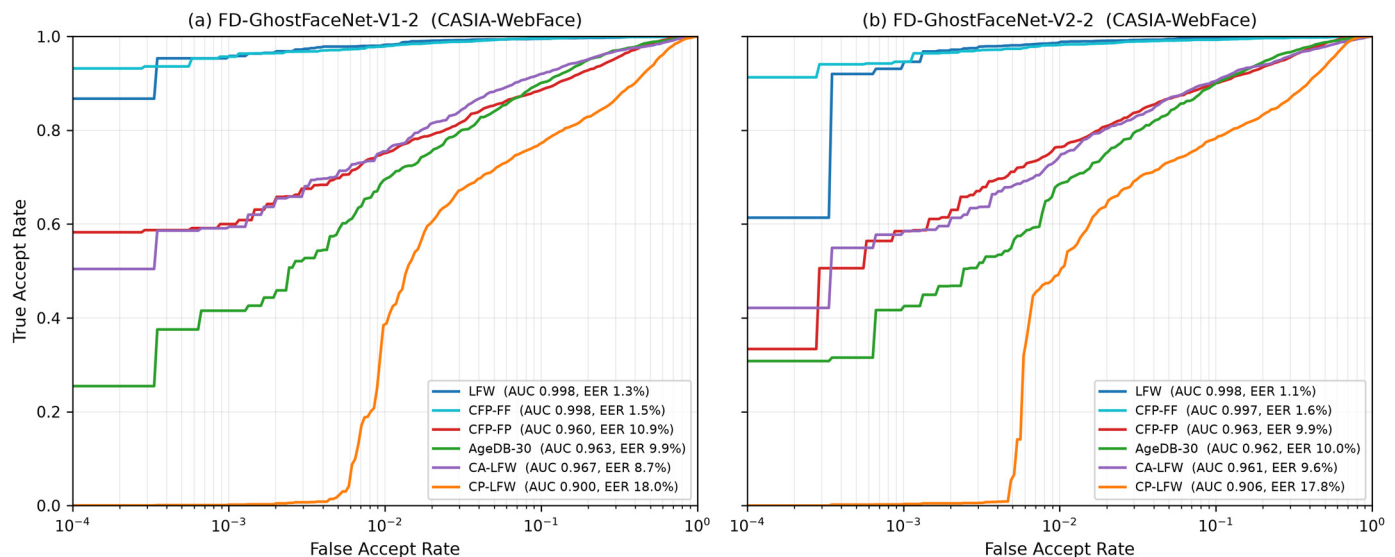


Figure 4. FD-GhostFaceNet ROC on CASIA-WebFace dataset.

Figure 5 traces verification accuracy on the six benchmarks across the 50-epoch schedule for FD-GhostFaceNet-V1-2 (a) and FD-GhostFaceNet-V2-2 (b). In both panels, accuracy rises sharply in the first few epochs and then stratifies by difficulty in the same order seen in the ROC analysis: the frontal benchmarks (LFW, CFP-FF) saturate above 98% within the first third of the schedule, the cross-age and frontal-to-profile benchmarks (CA-LFW, AgeDB-30, CFP-FP) form a closely spaced cluster climbing into the low 90s, and CP-LFW remains clearly below, plateauing near 84%. A transient dip is visible on the pose benchmarks around epochs 8–10 in both panels—most pronounced on CFP-FP—coinciding with the ArcFace margin becoming fully binding at the peak learning rate; both models recover within roughly two epochs to a strictly higher level, and all reported accuracies are taken from the converged region. Thereafter, every curve reaches a stable plateau without late-stage degradation, indicating clean convergence rather than overfitting. The two trunks behave consistently, with V2-2 passing through the transient more smoothly and ending marginally higher on the pose benchmarks, while CA-LFW marginally favors V1-2—mirroring the ROC comparison. The contrast between the early saturation of the frontal sets and the sustained late-training gains on the pose and age benchmarks is consistent with the model directing its additional capacity toward these harder conditions.

4.3.2. Performance on UMDFaces

To assess generalization across training corpora, the same two architectures are trained on UMDFaces with the unified recipe of Section 4.2. Table 4 summarizes the results. Both FD models improve on their same-trunk baselines across all six benchmarks—FD-GhostFaceNet-V1-2 by +0.05 to +2.97 points and FD-GhostFaceNet-V2-2 by +0.10 to +2.94 points—and the cross-pose gains are again the largest, setting new bests for the table: CFP-FP rises from 87.60% (ECA-CSP GhostFaceNet V2-2) to 89.56% (+1.96, FD-V1-2) and CP-LFW from 84.65% (CSP GhostFaceNet V2-2) to 86.12% (+1.47, FD-V2-2), with both FD models individually exceeding the previous best on both benchmarks; notably, the

lighter V1-2 trunk holds the CFP-FP record. On the frontal and cross-age benchmarks, the FD models land within the spread of the attention-augmented variants rather than above it: LFW reaches 98.92% against 99.08% for ECA-CSP V2-2, CA-LFW reaches 92.57% against 92.77%, CFP-FF reaches 98.61% against 98.71%, and AgeDB-30 reaches 91.52% against 92.13%. This profile matches the CASIA-WebFace results: the frequency partition buys its largest margins where identity must be recovered under pose change, while the near-saturated frontal sets and the cross-age benchmarks—the regime the channel-attention variants target—remain within a few tenths of a point. The consistency of this pattern across two training corpora of different scale and composition indicates that the gains stem from the architectural prior rather than from tuning to a particular dataset, and the concentration of the largest margins on the cross-pose benchmarks matches the design intent of the frequency-decoupled generator.

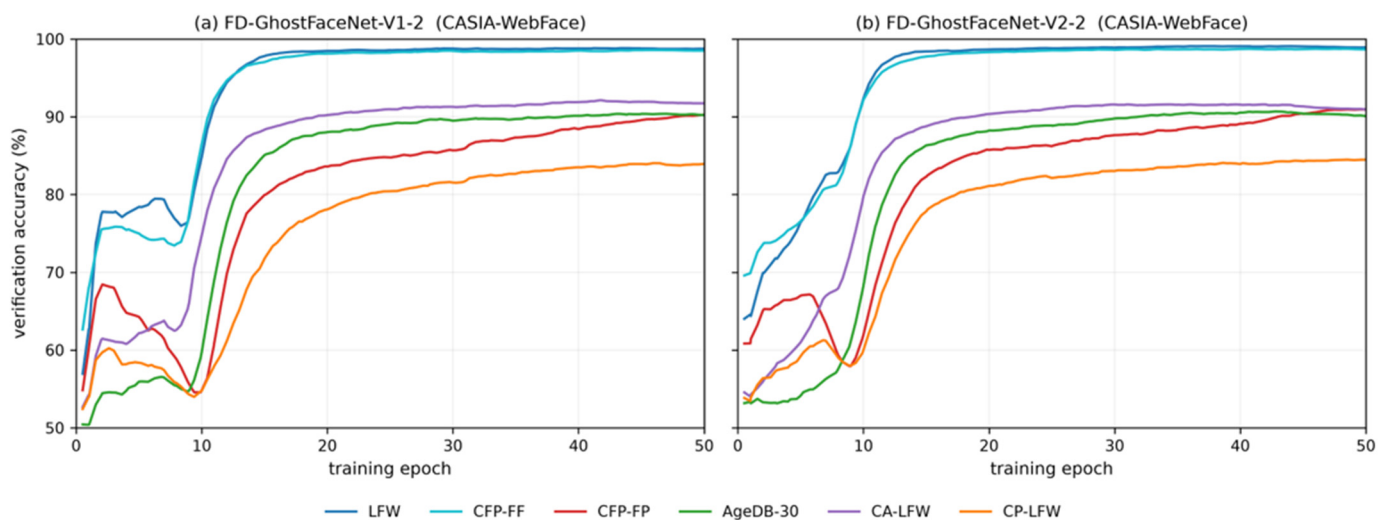


Figure 5. Verification accuracy on CASIA-WebFace dataset.

Table 4. Performance of models trained on the UMDFaces dataset; baseline and GhostFaceNet++ results.

Model	LFW (%)	CP-LFW (%)	CA-LFW (%)	AgeDB-30 (%)	CFP-FP (%)	CFP-FF (%)
GhostFaceNet V1-2 (Baseline)	98.77	83.72	92.20	90.62	86.59	98.41
CSP GhostFaceNet V1-2	98.82	83.45	92.13	90.70	86.93	98.40
ECA-CSP GhostFaceNet V1-2	98.77	83.60	92.27	90.98	87.08	98.11
FD-GhostFaceNet-V1-2 (Ours)	98.87	85.57	92.25	91.05	89.56	98.51
GhostFaceNet V2-2 (Baseline)	98.82	83.18	92.15	91.23	86.89	98.43
CSP GhostFaceNet V2-2	99.02	84.65	92.70	92.13	87.59	98.69
ECA-CSP GhostFaceNet V2-2	99.08	84.22	92.77	92.08	87.60	98.71
FD-GhostFaceNet-V2-2 (Ours)	98.92	86.12	92.57	91.52	89.31	98.61

All models are trained with the identical ArcFace configuration ($m = 0.5$, $s = 64$); bold rows denote the proposed models.

Figure 6 presents the ROC curves of FD-GhostFaceNet-V1-2 (a) and FD-GhostFaceNet-V2-2 (b) on the six UMDFaces benchmarks, evaluated from the final checkpoint of each model and plotted on a logarithmic FAR axis. Both panels reproduce the tiering seen on CASIA-WebFace, with one difference in the middle of the ranking. The near-frontal bench-

marks are saturated, with LFW and CFP-FF at AUCs of 0.997–0.998 (EER 1.5–1.7%); here, however, the cross-age benchmarks separate clearly from the frontal-to-profile one, with CA-LFW and AgeDB-30 forming the intermediate tier (AUC 0.967–0.970, EER 8.4–9.4%), while both pose benchmarks sit below it: CFP-FP at AUC 0.950–0.952 (EER 11.5–12.0%) and the extreme cross-pose CP-LFW at AUC 0.914–0.915 (EER 16.0–16.1%), whose true-acceptance rate falls steeply as FAR decreases. The two trunks are nearly indistinguishable in ROC terms—AUCs agree to within 0.003 on every benchmark—with V1-2 marginally ahead on CFP-FP, consistent with Table 4. Relative to the CASIA-WebFace models (Figure 4), the extreme cross-pose EER improves from $\approx 18\%$ to $\approx 16\%$, consistent with the heavier pose variation of UMDFaces, yet the pose pair still dominates the residual error on both trunks—confirming cross-pose verification as the binding constraint the global low-frequency pathway is designed to address.

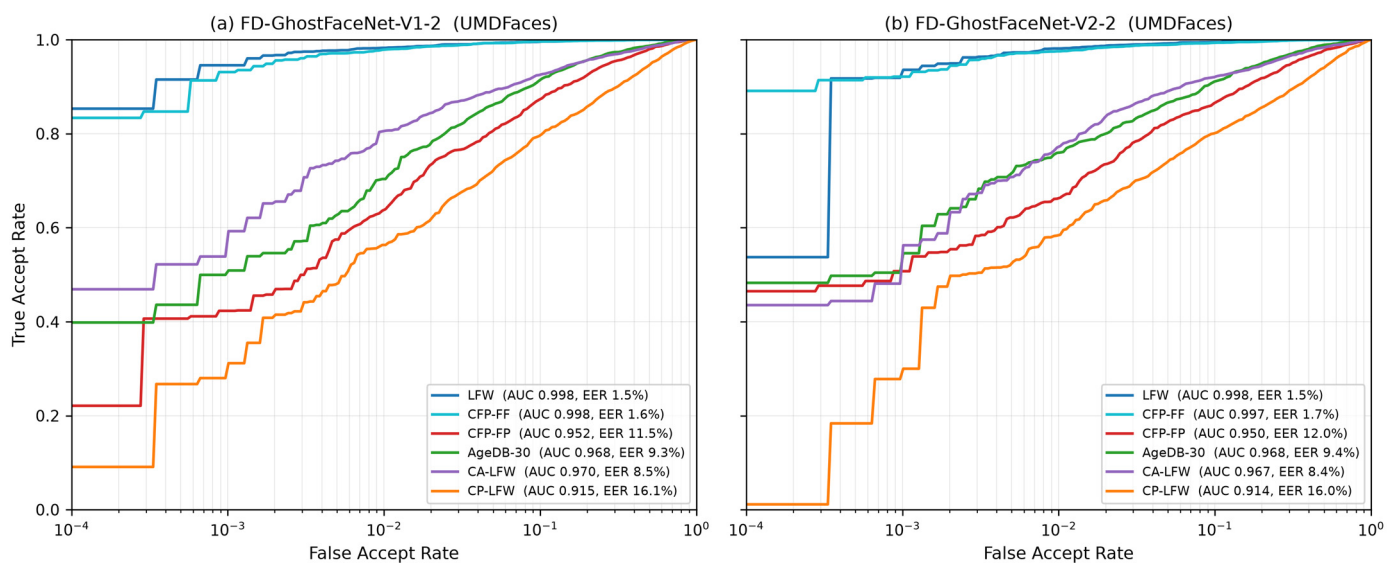


Figure 6. FD-GhostFaceNet ROC on UMDFaces dataset.

Figure 7 traces verification accuracy on the six benchmarks across the 50-epoch schedule for FD-GhostFaceNet-V1-2 (a) and FD-GhostFaceNet-V2-2 (b). In both panels accuracy rises sharply over the first few epochs and then stratifies by difficulty in the same order seen in the ROC analysis: the frontal benchmarks (LFW, CFP-FF) saturate above 98% within the first twenty epochs, the cross-age benchmarks (CA-LFW, AgeDB-30) climb steadily into the low 90s, and the pose benchmarks (CFP-FP, CP-LFW) remain clearly below, plateauing in the mid-to-high 80s. In contrast to the CASIA-WebFace runs (Figure 5), no transient dip appears near the peak learning rate—on the pose-richer UMDFaces corpus, both trunks progress essentially monotonically. All curves hold a stable plateau to the end of the schedule without late-stage degradation, indicating clean convergence rather than overfitting, with the sustained late-training gains concentrated on the harder pose and age benchmarks. The two panels track each other closely: V2-2 climbs somewhat faster through the first ten epochs and ends marginally higher on most benchmarks, while CFP-FP marginally favors V1-2, consistent with Table 4 and the ROC comparison.

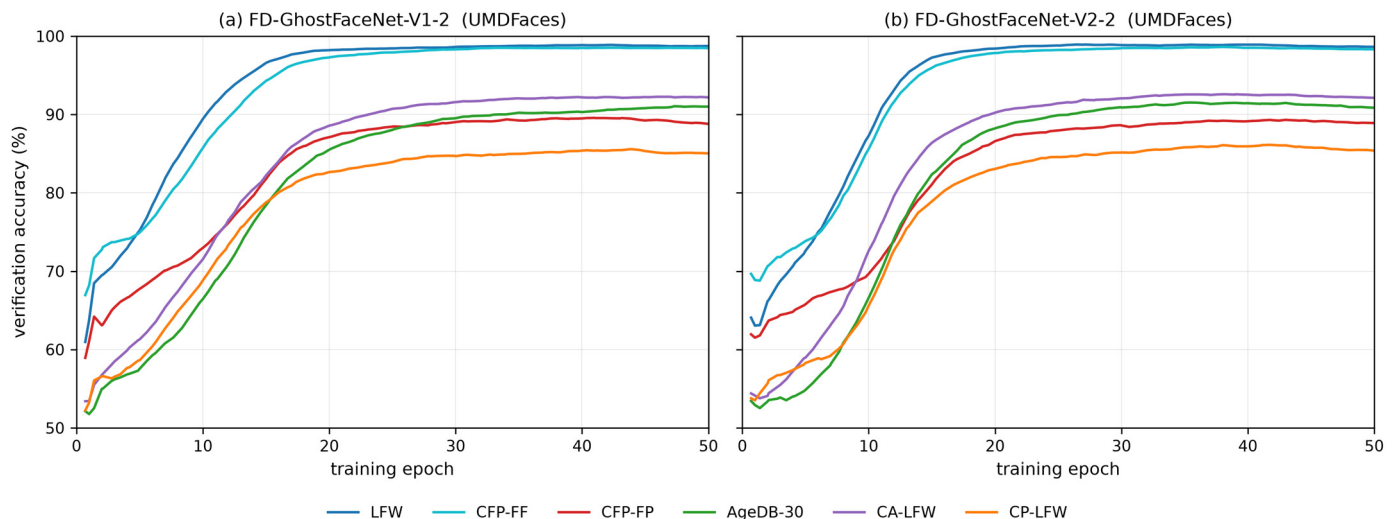


Figure 7. Verification accuracy on UMDFaces dataset.

4.3.3. Efficiency Analysis and Discussion

FD-GhostFaceNet-V1-2 operates at 83.1 MFLOPs, 4.60 M parameters, and 9.19 MB of fp16 storage, with FD-GhostFaceNet-V2-2 at 97.1 MFLOPs, 7.39 M parameters, and 14.77 MB. Measured under the same counting convention, the frequency-decoupled components add 0.53 M and 0.55 M parameters (+13.0% and +8.0%) and 22.8 and 20.6 MFLOPs to the GhostFaceNet V1-2 and V2-2 trunks (60.3 and 76.5 MFLOPs; 4.07 M and 6.84 M parameters; Table 3), and the two models exceed their same-trunk baselines in 23 of the 24 trunk–benchmark comparisons across the two training corpora, with the single exception (AgeDB-30, V2-2 on CASIA-WebFace) within 0.1 point. Relative to the strongest GhostFaceNet++ variants (51.6–62.0 MFLOPs), the two models spend an additional 21–46 MFLOPs and convert it into gains concentrated exactly where the existing lineage is weakest: on UMDFaces, CFP-FP rises by 1.96 points and CP-LFW by 1.47 points over the previous best, and on CASIA-WebFace, CFP-FP and LFW reach new bests, while the near-saturated frontal and the cross-age benchmarks remain within 0.1–0.6 points of the strongest attention variants. The models therefore occupy a deliberate point on the accuracy–efficiency frontier: they remain within the sub-100 MFLOPs regime, 4.5–5.3× below MobileFaceNet’s 439.8 MFLOPs, and buy a pose robustness that the cheaper variants do not reach. The additional computation is spent on the DCT-based global generators at the stages operating at 28×28 and below—with their input-conditioned gating, DC anchoring, and cross-band coupling—and on the anti-aliased downsampling, and the benchmark profile of the gains aligns with this design: the largest improvements arise precisely on the cross-pose benchmarks targeted by the global low-frequency pathway and the shift-invariant downsampling, while the frontal benchmarks are preserved at saturation. All efficiency figures are analytical multiply–accumulate counts under the reporting convention shared by the compared models; measured latency additionally depends on the maturity of spectral matrix-multiplication kernels in the target runtime, and hardware-specific profiling is left to future work.

Two further observations support the robustness of the result. First, although the tabulated accuracies follow the per-benchmark reporting convention of the baselines’ framework, the ROC analysis is produced from a single final checkpoint per model, and those operating points sustain the same benchmark profile—a single set of weights delivers the reported behavior across all six benchmarks. Second, the two architectures are identical across the training corpora, and the training objective, augmentation, schedule, and evaluation protocol are matched to those of the baselines; the comparison therefore primarily

isolates the backbone, in which the proposed frequency-decoupled components are the changed variables. These results indicate that, within the sub-100 MFLOPs regime, allocating a lightweight model's capacity to an explicit global low-frequency identity pathway is a more effective architectural prior to pose-robust verification than the purely local feature generation—or the channel-attention refinements—of the existing Ghost lineage.

4.3.4. Positioning Among State-of-the-Art Lightweight Models

Table 5 situates the FD-GhostFaceNet models within the broader landscape of lightweight face recognition architectures. Models trained on the MS-Celeb-1 M variants benefit from roughly an order-of-magnitude more training images (5.1 M–5.8 M versus 0.49 M and 0.37 M) and are included for context rather than direct comparison; under such large-scale training, the GhostFaceNet family itself reports LFW accuracies up to 99.85%.

Table 5. Comparison with state-of-the-art lightweight face recognition models. A dash indicates a value not reported in the source. Bold rows denote the proposed models, and the bold spanning rows group the entries by training dataset.

Model	FLOPs (M)	Params (M)	Size (MB)	LFW (%)
Trained on UMDFaces (0.37 M)				
ProxylessFaceNAS [36]	873.95	3.01	12.40	98.82
ECA-CSP GhostFaceNet V1-2 [10]	51.65	3.75	7.51	98.77
ConvFaceNeXt_TE [37]	410.59	1.05	4.49	99.05
CSP GhostFaceNet V1-2 [10]	51.64	3.75	7.51	98.82
ConvFaceNeXt_PE [37]	404.57	1.05	4.49	99.10
CSP GhostFaceNet V2-2 [10]	61.98	5.83	11.66	99.02
ConvFaceNeXt_TS [37]	396.15	0.96	4.12	99.17
MobileFaceNet [36]	439.8	1.03	4.50	99.03
ECA-CSP GhostFaceNet V2-2 [10]	62.00	5.83	11.66	99.08
FD-GhostFaceNet-V1-2 (ours)	83.10	4.60	9.19	98.87
FD-GhostFaceNet-V2-2 (ours)	97.08	7.39	14.77	98.92
Trained on CASIA-WebFace (0.49 M)				
ECA-CSP GhostFaceNet V1-2 [10]	51.65	3.75	7.51	98.87
1.00 Fast-FaceNet [38]	684.00	4.80	—	98.15
LFR (MobileFaceNet + ECA) [39]	—	—	3.40	98.52
CSP GhostFaceNet V2-2 [10]	61.98	5.83	11.66	99.03
0.75 Fast-FaceNet [38]	392.00	3.10	—	97.25
ECA-CSP GhostFaceNet V2-2 [10]	62.00	5.83	11.66	98.88
CSP GhostFaceNet V1-2 [10]	51.64	3.75	7.51	98.87
CNN [21]	5015.00	—	—	97.73
FD-GhostFaceNet-V1-2 (ours)	83.10	4.60	9.19	98.82
FD-GhostFaceNet-V2-2 (ours)	97.08	7.39	14.77	99.08

Among models trained on CASIA-WebFace, FD-GhostFaceNet-V2-2 attains 99.08% on LFW at 97.1 MFLOPs and 14.77 MB, the highest accuracy of any Ghost-based model in this group, with FD-GhostFaceNet-V1-2 reaching 98.82% at 83.1 MFLOPs. MobileFaceNets and AirFace report marginally higher LFW accuracies (99.28% and 99.27%) but require approximately 4.5 and 10 times the computation, respectively, while the Fast-FaceNet variants demand 4 to 7 times the computation at substantially lower accuracy, and LFR, though compact (3.40 MB), trails by 0.56 percentage points. Among models trained on UMDFaces, the FD models reach 98.92% and 98.87%; the ConvFaceNeXt variants report up to 99.17% but require four to five times the computation, and none of the higher-scoring entries in this group reports the cross-pose benchmarks on which the FD models set the family's records (Table 4).

It should be noted that LFW is a near-saturated, predominantly frontal benchmark and is reported here because it is the only benchmark shared across all listed methods; the distinguishing strengths of FD-GhostFaceNet lie on the cross-pose and cross-age benchmarks of Tables 3 and 4, which most lightweight architectures do not report under these small-scale training protocols. Taken together with those results, the comparison shows that the FD models offer a competitive accuracy-to-computation trade-off among lightweight models trained on small-scale data, with their advantage concentrated precisely on the hardest verification conditions.

5. Conclusions

This paper introduced FD-GhostFaceNet, a lightweight face recognition backbone built upon the proposed Frequency-Decoupled Ghost (FD-Ghost) module. FD-Ghost replaces the strictly local, cheap generator of the Ghost module at all stages operating at 28×28 and below with a complementary-band global counterpart: a fixed orthonormal two-dimensional DCT, partitioned by a learnable per-channel, per-frequency gate with an input-conditioned shift of the band boundary, decomposes the intrinsic features into a global low-frequency band biased toward pose- and age-stable identity structure and a locally refined high-frequency detail band, with exact reconstruction guaranteeing no representational loss at initialization. DC anchoring routes the most stable spectral component unconditionally through the low band, a bounded cross-band bilinear interaction contributes second-order capacity without additional depth, and the architecture is completed by anti-aliased downsampling at the deep stride sites. Trained with the unchanged ArcFace recipe of the GhostFaceNet++ protocol on both trunks of the family, FD-GhostFaceNet-V1-2 operates at 4.60 M parameters, 83.1 MFLOPs, and 9.19 MB of fp16 storage, with FD-GhostFaceNet-V2-2 at 7.39 M parameters, 97.1 MFLOPs, and 14.77 MB; both are within the sub-100 MFLOPs regime, and both exceed their same-trunk baselines in 23 of the 24 trunk-benchmark comparisons across the two training corpora, with the single exception (AgeDB-30, V2-2 on CASIA-WebFace) within 0.1 point. The gains concentrate where the existing lineage is weakest: trained on UMDFaces, the FD models raise the best reported CP-LFW accuracy from 84.65% to 86.12% and CFP-FP from 87.60% to 89.56%; trained on CASIA-WebFace, they improve CFP-FP from 90.10% to 90.94% and set a new best LFW accuracy of 99.08%, while the remaining frontal and cross-age benchmarks stay within a few tenths of the strongest attention-augmented variants. This concentration of the largest gains on the cross-pose benchmarks matches the design intent: the global low-frequency pathway and shift-invariant downsampling target precisely the conditions under which purely local feature generation fails. These results confirm that, within a fixed lightweight budget, an explicit frequency-decoupled global pathway is a more effective architectural prior for pose-robust verification than the purely local synthesis of the existing Ghost lineage, and its cost profile of sub-100 MFLOPs and at most 14.8 MB of fp16 storage makes both models strong candidates for deployment on resource-constrained devices. All comparative claims in this paper are accordingly scoped to the small-scale CASIA-WebFace and UMDFaces training protocols under which the baseline results are reported. Future work will examine whether the benefits of frequency-decoupled generation persist under data-rich training on large-scale corpora such as Glint360K and will extend the approach to template-based benchmarks and other lightweight backbones, together with latency profiling on dedicated mobile inference runtimes.

Author Contributions: Conceptualization, A.Q.; Methodology, A.Q.; Formal Analysis, A.Q.; Writing—Original Draft Preparation, A.Q.; Visualization, A.Q.; Supervision, B.A.H. and H.K.; Validation, B.A.H. and H.K.; Investigation, B.A.H. and H.K.; Writing—Review and Editing, H.K.; Formal Analysis, H.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

- Deng, J.; Guo, J.; Xue, N.; Zafeiriou, S. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2019; pp. 4690–4699.
- Wang, H.; Wang, Y.; Zhou, Z.; Ji, X.; Gong, D.; Zhou, J.; Li, Z.; Liu, W. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2018; pp. 5265–5274.
- Huang, G.B.; Mattar, M.; Berg, T.; Learned-Miller, E. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In *Proceedings of the Workshop on faces in 'Real-Life' Images: Detection, Alignment, and Recognition*, Marseille, France, 17 October 2008.
- Chen, S.; Liu, Y.; Gao, X.; Han, Z. Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices. In *Proceedings of the Chinese Conference on Biometric Recognition*; Springer International Publishing: Cham, Switzerland, 2018; pp. 428–438.
- Martinez-Diaz, Y.; Luevano, L.S.; Mendez-Vazquez, H.; Nicolas-Diaz, M.; Chang, L.; Gonzalez-Mendoza, M. Shufflefacenet: A lightweight face architecture for efficient and highly-accurate face recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*; IEEE: New York, NY, USA, 2019; pp. 2721–2728.
- Yan, M.; Zhao, M.; Xu, Z.; Zhang, Q.; Wang, G.; Su, Z. Vargfacenet: An efficient variable group convolutional neural network for lightweight face recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*; IEEE: New York, NY, USA, 2019; pp. 2647–2654.
- Han, K.; Wang, Y.; Tian, Q.; Guo, J.; Xu, C.; Xu, C. Ghostnet: More features from cheap operations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2020; pp. 1580–1589.
- Tang, Y.; Han, K.; Guo, J.; Xu, C.; Xu, C.; Wang, Y. GhostNetv2: Enhance cheap operation with long-range attention. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 9969–9982. [\[CrossRef\]](#)
- Alansari, M.; Hay, O.A.; Javed, S.; Shoufan, A.; Zweiri, Y.; Werghi, N. Ghostfacenets: Lightweight face recognition model from cheap operations. *IEEE Access* **2023**, *11*, 35429–35446. [\[CrossRef\]](#)
- Nachet, R.; Garrigós, J.; Stambouli, T.B. GhostFaceNet++: Boosting efficiency and accuracy via CSP bottlenecks and Channel Attention. *J. Real-Time Image Process.* **2025**, *22*, 201. [\[CrossRef\]](#)
- Zheng, T.; Deng, W. Cross-pose lfw: A database for studying cross-pose face recognition in unconstrained environments. *Beijing Univ. Posts Telecommun. Tech. Rep.* **2018**, *5*, 5.
- Sengupta, S.; Chen, J.-C.; Castillo, C.; Patel, V.M.; Chellappa, R.; Jacobs, D.W. Frontal to profile face verification in the wild. In *Proceedings of the 2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*; IEEE: New York, NY, USA, 2016; pp. 1–9.
- Zheng, T.; Deng, W.; Hu, J. Cross-age lfw: A database for studying cross-age face recognition in unconstrained environments. *arXiv* **2017**, arXiv:1708.08197.
- Moschoglou, S.; Papaioannou, A.; Sagonas, C.; Deng, J.; Kotsia, I.; Zafeiriou, S. Agedb: The first manually collected, in-the-wild age database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*; IEEE: New York, NY, USA, 2017; pp. 51–59.
- Ahmed, N.; Natarajan, T.; Rao, K.R. Discrete cosine transform. *IEEE Trans. Comput.* **1974**, *100*, 90–93. [\[CrossRef\]](#)
- Qin, Z.; Zhang, P.; Wu, F.; Li, X. Fcanet: Frequency channel attention networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*; IEEE: New York, NY, USA, 2021; pp. 783–792.
- Rao, Y.; Zhao, W.; Zhu, Z.; Lu, J.; Zhou, J. Global filter networks for image classification. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 980–993.
- Chen, Y.; Fan, H.; Xu, B.; Yan, Z.; Kalantidis, Y.; Rohrbach, M.; Yan, S.; Feng, J. Drop an octave: Reducing spatial redundancy in convolutional neural networks with octave convolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*; IEEE: New York, NY, USA, 2019; pp. 3435–3444.
- Zhang, R. Making convolutional networks shift-invariant again. In *Proceedings of the International Conference on Machine Learning*, Long Beach, CA, USA, 10–15 June 2019; pp. 7324–7334.
- Bansal, A.; Nanduri, A.; Castillo, C.D.; Ranjan, R.; Chellappa, R. Umdfaces: An annotated face dataset for training deep networks. In *Proceedings of the 2017 IEEE International Joint Conference on Biometrics (IJCB)*; IEEE: New York, NY, USA, 2017; pp. 464–473.
- Yi, D.; Lei, Z.; Liao, S.; Li, S.Z. Learning face representation from scratch. *arXiv* **2014**, arXiv:1411.7923.
- Iandola, F.N.; Han, S.; Moskewicz, M.W.; Ashraf, K.; Dally, W.J.; Keutzer, K. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size. *arXiv* **2016**, arXiv:1602.07360.

23. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.
24. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.-C. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2018; pp. 4510–4520.
25. Zhang, X.; Zhou, X.; Lin, M.; Sun, J. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2018; pp. 6848–6856.
26. Ma, N.; Zhang, X.; Zheng, H.-T.; Sun, J. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European Conference on Computer Vision (ECCV)*; IEEE: New York, NY, USA, 2018; pp. 116–131.
27. Boutros, F.; Damer, N.; Fang, M.; Kirchbuchner, F.; Kuijper, A. Mixfacenets: Extremely efficient face recognition networks. In *Proceedings of the 2021 IEEE International Joint Conference on Biometrics (IJCB)*; IEEE: New York, NY, USA, 2021; pp. 1–8.
28. Boutros, F.; Siebke, P.; Klemt, M.; Damer, N.; Kirchbuchner, F.; Kuijper, A. Pocketnet: Extreme lightweight face recognition network using neural architecture search and multistep knowledge distillation. *IEEE Access* **2022**, *10*, 46823–46833. [[CrossRef](#)]
29. George, A.; Ecabert, C.; Shahreza, H.O.; Kotwal, K.; Marcel, S. Edgeface: Efficient face recognition model for edge devices. *IEEE Trans. Biom. Behav. Identity Sci.* **2024**, *6*, 158–168. [[CrossRef](#)]
30. Liu, Z.; Hao, Z.; Han, K.; Tang, Y.; Wang, Y. Ghostnetv3: Exploring the training strategies for compact models. *arXiv* **2024**, arXiv:2404.11202.
31. Guibas, J.; Mardani, M.; Li, Z.; Tao, A.; Anandkumar, A.; Catanzaro, B. Adaptive fourier neural operators: Efficient token mixers for transformers. *arXiv* **2021**, arXiv:2111.13587.
32. Gueguen, L.; Sergeev, A.; Kadlec, B.; Liu, R.; Yosinski, J. Faster neural networks straight from jpeg. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 3933–3944.
33. Li, Q.; Shen, L.; Guo, S.; Lai, Z. Wavelet integrated CNNs for noise-robust image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2020; pp. 7245–7254.
34. Qian, Y.; Yin, G.; Sheng, L.; Chen, Z.; Shao, J. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020*; pp. 86–103.
35. Li, J.; Xie, H.; Li, J.; Wang, Z.; Zhang, Y. Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2021; pp. 6458–6467.
36. Hoo, S.C.; Ibrahim, H.; Suandi, S.A.; Ng, T.F. Lcam: Low-complexity attention module for lightweight face recognition networks. *Mathematics* **2023**, *11*, 1694. [[CrossRef](#)]
37. Hoo, S.C.; Ibrahim, H.; Suandi, S.A. Convfacenext: Lightweight networks for face recognition. *Mathematics* **2022**, *10*, 3592. [[CrossRef](#)]
38. Xu, X.; Du, M.; Guo, H.; Chang, J.; Zhao, X. Lightweight facenet based on mobilenet. *Int. J. Intell. Sci.* **2021**, *11*, 1–16. [[CrossRef](#)]
39. Xiao, J.; Jiang, G.; Liu, H. A lightweight face recognition model based on mobilefacenet for limited computation environment. *EAI Endorsed Trans. Internet Things* **2022**, *7*, 1–9. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.