Fusion of Edge-Computing Technology for Smart-Security Face-Recognition System Development

Yangjing Cheng
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, Shaanxi, China
E-mail: 1431807602@qq.com

Rui Cao
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, Shaanxi, China
E-mail: 1275280817@qq.com

Abstract—**Traditional centralized face recognition systems suffer five critical engineering drawbacks: high network transmission latency, high risk of facial biometric privacy leakage, excessive cloud hardware and operation-maintenance costs, service failure under network disconnection, and poor recognition robustness under complex illumination and occlusion conditions. In this paper, MobileFaceNet is taken as the backbone network. An ECA efficient one-dimensional channel attention module is embedded to construct a CNN-ViT hybrid feature extraction network. Structured channel pruning combined with 8-bit layered mixed quantization is adopted to realize model lightweight deployment. A three-level distributed collaborative system named "edge terminal-edge node-cloud management platform" is built. Core inference tasks including facial feature extraction and identity matching are offloaded to local edge-device execution. Multi-dimensional tests are carried out based on public LFW face dataset and self-built security-oriented occlusion dataset containing masked faces, sunglasses occlusions, large-angle side-faces and backlight samples. Complete ablation experiments are designed to verify independent gain of each optimized module. Horizontal comparative evaluation is performed against multiple mainstream lightweight face models. Measured experimental results show that the lightweight model achieves 98.7 % recognition accuracy under non-occlusion conditions and maintains over 85 % accuracy under mixed-occlusion scenarios. The model parameter volume is reduced by 32 %, floating-point computation cost drops by 35 %, and storage footprint is compressed by 75 %. Single-frame inference latency is controlled within 50 ms, yielding a 2.8-fold speedup on edge devices. A single edge node supports concurrent processing of 48 video streams. Full recognition and local data storage can operate offline without network connection. Compared with conventional cloud-centric schemes, cloud-side hardware investment is cut by 70 %, bandwidth consumption decreases by 80 %, and long-term operational cost falls by 65 %. The proposed algorithm and system jointly balance recognition accuracy, real-time inference throughput and embedded-hardware resource overhead. It can be deployed in residential compounds, industrial parks, transportation hubs and other security scenarios, and provides a reproducible, engineering-ready optimization paradigm for edge-oriented lightweight face-recognition systems.**

Keywords-Edge Computing; Smart Security; Face Recognition; Mobilefacenet; ECA Attention; CNN-Vit; Model Lightweighting; Ablation Experiment

I. INTRODUCTION

A. Research Background and Significance

With large-scale deployment of Internet-of-Things, embedded hardware and deep-learning visual algorithms, face recognition with contact-free acquisition and high discriminability has become a core underlying technique for intelligent security applications including access control verification, campus surveillance and station security screening [17]. Conventional face-recognition solutions adopt centralized "front-end capture plus cloud computation" architecture. All surveillance video frames and facial images are uploaded to cloud servers for processing. Practical field deployments

reveal non-negligible pain points. Massive concurrent video-stream transmission causes network congestion, and recognition response latency fails to satisfy real-time early-warning requirements. Facial data belongs to highly sensitive biometric information; cross-network transmission brings risks of data theft and tampering. Large-scale high-performance server clusters must be maintained in the cloud, inducing high upfront hardware procurement and continuous datacenter operation costs. The whole service fully depends on network links; once network disconnection or cloud-server failure occurs, the recognition service stops completely. Moreover, pure-convolution lightweight models only extract local receptive-field features, and their accuracy degrades sharply for backlight-exposed, mask-occluded or heavily yaw-rotated face samples [18].

Edge computing, as a distributed computing paradigm, pushes computing resources and service logics toward data-generation terminals. Most data processing is completed locally on edge devices, which reduces cloud interaction frequency and mitigates inherent drawbacks of centralized architectures from system-architecture perspective [12]. National strategic documents including New Generation Artificial Intelligence Development Plan and Overall Layout Plan for Digital China explicitly promote deep integration between lightweight domestic security visual algorithms and edge hardware. Most existing edge-face-recognition research merely ports pre-trained models onto edge hardware without scenario-specific network optimization for real-world occluded security samples. Lightweight approaches are oversimplified, and algorithm designs lack deep fusion with edge-system architectures [19]. Against this background, this work implements both three-tier edge-computing system and ECA-CNN-ViT hybrid lightweight face algorithm. Technical bottlenecks are resolved from three dimensions: system architecture, feature-extraction network and model compression. The presented work delivers both theoretical innovation values and large-scale engineering-deployment significance.

B. Research Objectives and Core Innovations

1) Research Objectives:

Construct a three-tier distributed collaborative architecture: edge terminal-edge node-cloud management platform. Core face-recognition inference tasks are offloaded to edge-side local execution to relieve cloud computing burden, reduce network dependency and improve offline operational reliability. Build the ECA-CNN-ViT hybrid feature extraction architecture based on MobileFaceNet backbone. Lightweight attention and Transformer-based global modeling branches are introduced to remedy poor robustness of pure-convolution networks under occlusion and backlight conditions. Propose a collaborative lightweight solution combining structured pruning and layered mixed quantization. Overall model accuracy loss is strictly constrained below 1 % while parameters, computation complexity and storage footprint are greatly reduced for low-resource edge gateways. Establish a hierarchical and standardized complete experimental validation framework. Ablation experiments quantify individual performance contributions of each improved module. Horizontal comparisons against mainstream lightweight models are performed under unified experimental settings, together with multi-scenario visual testing to comprehensively verify algorithm and system performance.

2) Core Innovations:

a) ECA-embedded CNN-ViT hybrid facial-feature extraction architecture:

MobileFaceNet inverted residual convolutions extract local facial texture. A single-layer lightweight linear ViT branch is embedded at high-level network layers for global facial pixel correlation modeling. The ECA one-dimensional channel-attention module adaptively amplifies weights of high-discriminative regions such as eyes and mouth, and suppresses interference introduced by masks and background noise. Local texture and global semantic features are mutually complemented. Occlusion-scenario accuracy increases by 9.9 % compared with vanilla MobileFaceNet [5], [7].

b) Customized joint lightweight strategy for CNN-ViT hybrid networks:

Existing pruning-quantization pipelines are mainly designed for pure-CNN networks. This paper applies differentiated layer-wise processing for hybrid network components. Structured channel pruning is conducted over convolutional layers. Shallow layers adopt 8-bit integer quantization, whereas ECA attention units and feature-fusion layers retain 16-bit half-precision floating-point computation. Under the constraint of ≤1 % accuracy degradation, 75 % storage compression is achieved while preserving fine-grained facial feature information [6], [10].

c) Multi-dimensional comprehensive experimental validation framework oriented toward real-world security scenarios:

Instead of relying solely on non-occluded public datasets, both standard LFW benchmark and a self-built 50 000-sample multi-occlusion security dataset are adopted. Gradient-style ablation experiments independently verify contributions of ECA and CNN-ViT components. Cross-model comparisons are executed on real embedded hardware. Static-image, dynamic-video and real-time camera visual verifications are supplemented. The reproducible experimental workflow provides a standardized evaluation paradigm for similar lightweight face-recognition research.

II. Related Works

A. Face-Recognition Algorithm Research

Face-recognition techniques are divided into handcrafted-feature and deep-learning-based stages. Classical algorithms such as Eigenface and Fisherface adopt PCA and LDA for manual feature dimensionality reduction. They work stably only under uniform-illumination occlusion-free conditions and exhibit poor generalization for real-world security scenes [1], [2]. After 2014, deep convolutional networks represented by DeepFace and FaceNet achieved remarkable accuracy improvements, yet their huge parameter volumes limited deployment exclusively to high-performance cloud servers [3], [9].

Lightweight convolutional networks emerged for mobile and edge-device adaptation. MobileFaceNet proposed in 2018 optimized inverted-residual structures specifically for face tasks and became the mainstream embedded backbone. Nevertheless, pure-convolution networks have limited receptive fields and suffer severe performance degradation for occluded samples [10]. ShuffleFaceNet proposed in 2020 compressed parameters via channel shuffling but lacked global modeling capacity, further degrading accuracy under complex conditions [11]. Starting from 2020, CNN-ViT hybrid architectures became popular for image recognition, where convolutions capture local patterns and Transformers model global dependencies. Existing hybrid face models adopt standard multi-head attention with excessive computational overhead. They rarely integrate lightweight ECA attention for occlusion-noise suppression and cannot be deployed cost-effectively on edge hardware [4], [7].

B. Model-Lightweighting Technology Research

Four mainstream lightweighting approaches include structured pruning, numerical quantization, knowledge distillation and neural-architecture search. Since 2020, pruning-plus-quantization joint schemes have become the dominant choice for edge-oriented models. Structured pruning removes redundant convolutional channels completely and produces dense networks compatible with NPU acceleration. However, existing pruning criteria target pure-convolution networks; direct application to CNN-ViT hybrid networks damages global self-attention weight distributions [10]. Uniform 8-bit quantization tends to discard fine-grained facial features. Layered mixed quantization balances accuracy and compression ratio, yet few existing studies combine layered mixed quantization with structured channel pruning [12], [13]. Knowledge distillation and NAS demand massive training computation resources and are impractical for small-to-medium-size security-project iterations.

C. Edge-Computing Applications in Security-Oriented Face Recognition

Overseas vendors such as IBM and NTT have released edge-AI security platforms, but high

device-procurement and deployment costs restrict their application for domestic small-and-medium-sized industrial parks [14], [16]. Domestic researchers including Zhang Jun and Li Peng built edge-based face-recognition systems. However, they only accomplished simple model migration without occlusion-oriented network optimization. Lightweighting means remain simplistic, and algorithm-architecture integration is insufficient [17], [18]. Huawei edge-inference platforms provide general-purpose inference frameworks but lack customized algorithm optimization for face-occlusion scenarios [19]. Summarizing existing literature, three pervasive drawbacks are observed: isolation between algorithm and system architecture, missing network optimization for security-scene interferences, and incomplete experimental verification systems.

D. Summary of Research Status

Current lightweight edge-face-recognition research exhibits four major shortcomings. First, pure-convolution lightweight networks such as MobileFaceNet and ShuffleFaceNet lack global semantic modeling capability; accuracy drops sharply under mask occlusion, large-angle yaw rotation and backlight [3], [11]. Second, existing CNN-ViT hybrid face models impose heavy computation burdens. Without lightweight ECA attention modules, low-cost edge hardware cannot sustain full inference pipelines [4], [5]. Third, lightweighting usually applies single pruning or quantization. Layer-aware joint compression strategies tailored for convolution-Transformer hybrid networks are rare, making it difficult to constrain accuracy loss within engineering-acceptable ranges [10], [12]. Fourth, many publications test only on occlusion-free public datasets without ablation studies. Real-world occluded samples are seldom used for validation, which brings weak experimental persuasiveness.

Targeting the above gaps, this paper builds the ECA-CNN-ViT hybrid network based on MobileFaceNet, designs layered structured pruning plus mixed-quantization joint lightweighting, constructs the three-tier distributed edge system, and establishes dual-dataset multi-control-group experimental protocols to fill existing research gaps.

III. Core Algorithm And System Architecture Design

A. Overall System Architecture

A three-tier distributed collaborative architecture: edge terminal-edge node-cloud management platform is designed as the deployment carrier for the lightweight ECA-CNN-ViT algorithm. All core inference jobs including facial-feature extraction and identity matching run locally on edge nodes. The cloud platform undertakes only auxiliary management responsibilities. Layer-wise functional partitions are described below.

1) Edge-Terminal Layer:

Composed of 720P@30fps high-definition surveillance cameras. It captures facial videos and images, executes preprocessing steps including grayscale conversion, Gaussian noise reduction and face cropping. Only cropped face images are transmitted to edge nodes instead of raw high-bitrate video streams, which greatly reduces uplink bandwidth consumption.

2) Edge-Node Layer:

Deployed on embedded gateways with 2 TOPS NPU. The lightweight ECA-CNN-ViT model is loaded here to implement facial-feature extraction, Euclidean-distance-based identity matching and abnormal-behavior analysis. Local SQLite databases store facial templates and recognition logs encrypted by AES-256. Full-pipeline recognition and local caching can be completed under network-disconnection conditions. Incremental data synchronizes automatically toward cloud after network recovery.

3) Cloud-Management-Platform Layer:

Computationally-heavy inference workloads are eliminated. Functions include device registration and status monitoring, cloud-side backup of facial templates, remote unified lightweight-model pushing, visualized recognition-data statistics, and RBAC multi-role permission management.

B. Overall Workflow of ECA-CNN-ViT Algorithm

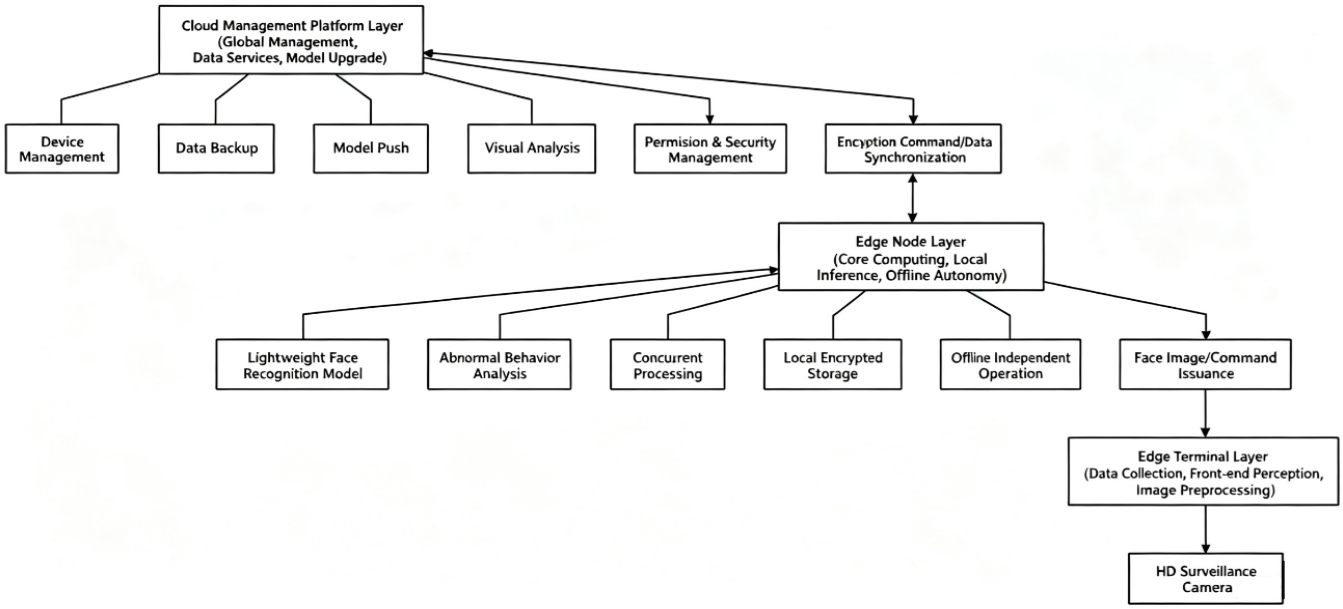

Figure 1. Three-tier distributed collaborative system architecture diagram.

The complete pipeline of the lightweight ECA-CNN-ViT algorithm consists of three modules: face preprocessing, hybrid feature extraction, and two-stage lightweight compression. Standardized 112×112 face images are input. After noise reduction and alignment preprocessing, shallow-layer MobileFaceNet inverted-residual convolutions extract local facial texture.

High-level feature maps split into two branches: one continues deep convolution processing, while the other feeds into the lightweight linear ViT branch for global-pixel-dependency modeling. Two-branch features are fused and fed into the ECA one-dimensional attention unit for adaptive high-value-feature screening. The network is supervised by ArcFace loss during training. After convergence, layered structured pruning, fine-tuning and layered mixed quantization are sequentially applied to produce the final edge-deployable lightweight inference model.

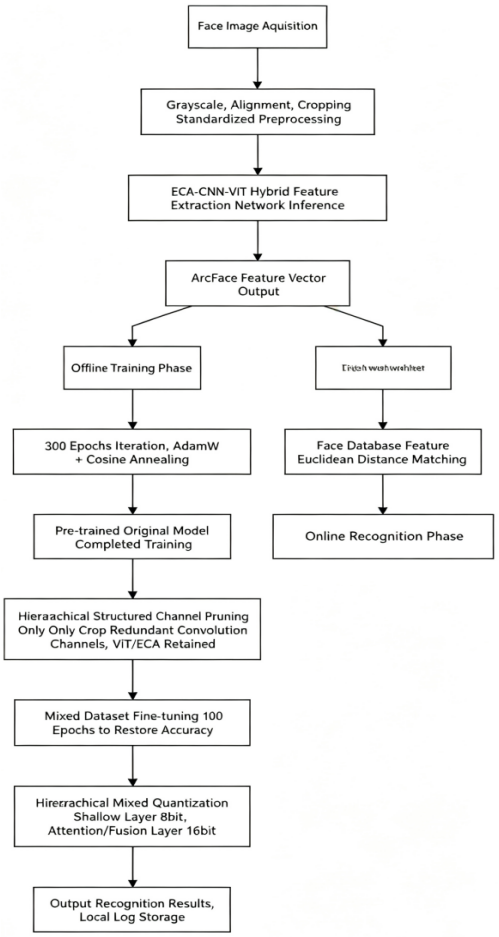

Figure 2. Workflow diagram of lightweight ECA-CNN-ViT face-recognition algorithm.

C. Base Backbone Network: MobileFaceNet

MobileFaceNet is built upon MobileNetV2 inverted-residual depth-separable convolutions. Channel configurations are optimized for 112×112 face input resolution. Its baseline accuracy reaches 97.2 % on the LFW non-occlusion dataset with parameter volume of merely 22.3 M, 80 % computation reduction compared with FaceNet. It serves as an optimal base backbone for edge-side face recognition [10]. Nevertheless, inherent local-receptive-field limitations of convolution prevent capturing long-range global dependencies among facial organs. Recognition accuracy drops to only 78.6 % for mask-occluded or over-45° yaw-rotated faces. Therefore, the ViT global branch and ECA attention are introduced for structural enhancement.

D. ECA Efficient One-Dimensional Channel-Attention Module

- ECA-Net abandons redundant dual-branch spatial-plus-channel computation adopted by CBAM. Channel correlation modeling is implemented purely via one-dimensional convolution with less than 8 % extra computational overhead, making it suitable for lightweight face networks [5]. The calculation flow is listed as follows.
- Global average pooling is performed over output feature maps to compress spatial dimensions and obtain one-dimensional channel-feature vectors.
- The one-dimensional convolution kernel size is adaptively computed according to total channel numbers to learn channel weights.

The Sigmoid function normalizes weights within the range 0~1. Learned weights multiply element-wise with original feature maps to amplify weights of facial-organ regions and suppress mask-and-background noise interference.

E. Lightweight Linear CNN-ViT Hybrid Feature Branch

Convolution excels at local-texture extraction, whereas ViT self-attention captures global-range facial dependencies. Standard multi-head ViT brings heavy computational overhead. This work adopts a single-layer lightweight linear multi-head self-attention ViT branch. Intermediate-level feature maps from MobileFaceNet are split into patches and transformed into token sequences. Position encoding is added for global-pixel interaction. Feature maps are restored and fused with convolution-branch outputs. Local details and global semantic information are jointly preserved, yielding obvious improvements for occluded samples with moderate computation overhead increment [7].

F. Structured Pruning plus Layered Mixed Quantization Lightweight Pipeline

1) Layer-Wise Structured Channel Pruning:

Only convolutional layers undergo channel pruning. ViT and ECA modules remain untouched. The absolute-value sum of convolutional-layer channel weights is counted layer-by-layer. Redundant channels whose weight sums fall below predefined thresholds are pruned. The overall pruning ratio is constrained below 30 %. After pruning, 100-epoch fine-tuning is conducted on mixed-occlusion datasets for accuracy recovery, reducing total parameters by 32 % [6].

2) Layered Mixed Quantization:

Shallow convolutional layers and ViT token-encoding layers adopt 8-bit integer compression. ECA attention units, feature-fusion layers and terminal fully-connected matching layers keep 16-bit half-precision floating-point format. KL-divergence calibration is introduced to mitigate quantization errors. Model storage volume is compressed by 75 %.

G. Loss Function and Training Strategy

ArcFace angular margin loss is adopted to enlarge inter-class feature distances and shrink intra-class distances among facial identities.

$$L_{ArcFace} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s(\cos(\theta_{y_i}+m))}}{e^{s(\cos(\theta_{y_i}+m))}+\sum_{j\neq y_i}e^{s\cos\theta_j}} \quad (1)$$

Equation (1) can be directly copied into MathType for rendering. Training

hyper-parameters: AdamW optimizer, initial learning rate 10^3, cosine-annealing learning-rate decay, total training epochs 300, batch size 32. Data-augmentation operations including random horizontal flip, brightness perturbation and random rotation are applied on training samples to mitigate overfitting risks.

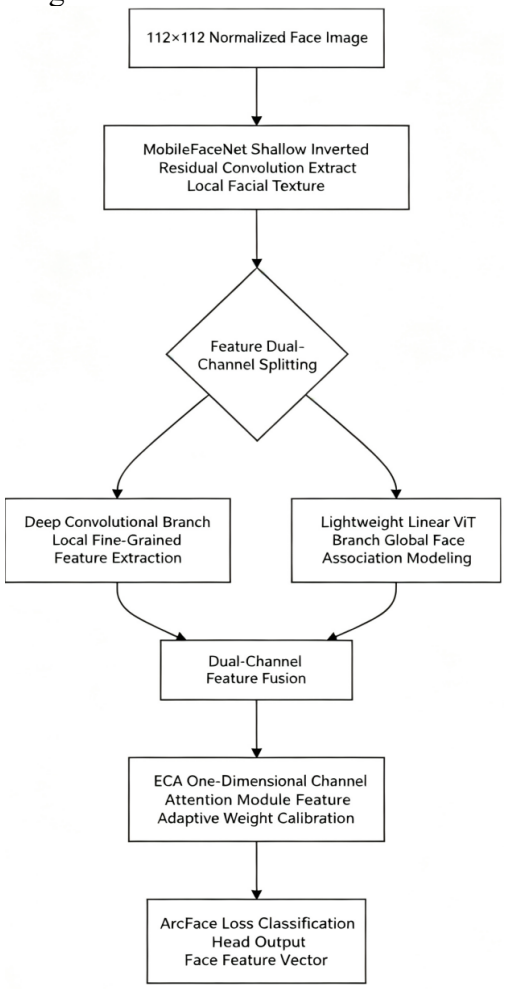

Figure 3. Network structure of ECA-CNN-ViT hybrid facial-feature extraction.

H. Quantitative Evaluation Metrics

- Accuracy metrics: non-occlusion top-1 accuracy Accclear, occlusion top-1 accuracy Accocclu.
- Real-time performance metrics: single-frame inference latency Lat (ms) on edge hardware.
- Hardware-overhead metrics: model parameter count (M), floating-point operations (G), model storage size (MB).
- Auxiliary classification metrics: precision, recall, F1-score.

IV. EXPERIMENTAL SETUP: HARDWARE-SOFTWARE ENVIRONMENT

A. Experimental Objectives

Lightweighting-control experiments measure changes of accuracy, parameter volume and inference latency induced by the pruning-plus-mixed-quantization pipeline, verify compression efficiency and accuracy controllability. Ablation experiments remove CNN-ViT and ECA modules sequentially to quantify independent accuracy gains contributed by each component. Cross-model horizontal comparison compares the proposed model against MobileFaceNet, ShuffleFaceNet and MobileNetV3-Face under unified hardware-software conditions. Multi-scenario visual validation includes static-image, dynamic-video and real-time camera testing to observe practical face-detection and identity-matching behaviors.

B. Hardware and Software Environment

TABLE I. EXPERIMENTAL HARDWARE CONFIGURATION

Device Type	Detailed Hardware Configuration	Experimental Purpose
Model-training workstation	Intel Core i7-10700K CPU, 16 GB RAM, GTX 1660 6 GB GPU	Network training, pruning fine-tuning, quantization testing, offline performance profiling
Edge-inference gateway	Quad-core ARM Cortex-A53, 2 TOPS embedded NPU, 4 GB RAM, 32 GB local storage, Ethernet /4G/5G support	Lightweight-model deployment, local face-recognition inference, offline identification and storage
Image-acquisition device	720P 30fps wide-dynamic-range infrared surveillance camera	Face-image and real-time-video capture for visual evaluation
Cloud-management server	Intel Xeon E5-2680 v4, 32 GB RAM, 1 TB SSD, gigabit Ethernet	Background-management-platform deployment, template backup, remote-device management, data aggregation

TABLE II. EXPERIMENTAL SOFTWARE CONFIGURATION

Device Carrier	Operating System	Core Development / Inference Framework	Programming Language	Database	Supporting Test Tools
Training workstation	Windows 11	PyTorch 1.10, TorchVision 0.11	Python 3.8	MySQL 8.0	JUnit5, OWASP ZAP
Edge gateway	Ubuntu 20.04 LTS embedded edition	PyTorch-Lite inference library	Python 3.8	SQLite 3	OpenCV, JMeter concurrency tester
Cloud server	Windows Server 2019	Django 4.0	Python 3.8	MySQL 8.0	Grafana, Prometheus monitoring toolkit

C. Datasets

A combination of public benchmark dataset and self-built security-scene dataset guarantees comprehensiveness and practical-scenario fitness. The LFW dataset is a world-widely-adopted public face-recognition benchmark containing 13 233 face images of 5 749 individuals with diverse gender, age and race distributions, mostly free of heavy occlusion, used for baseline accuracy evaluation. The self-built security-oriented dataset contains 50 000 images covering 5 000 identities (10 samples per identity). Samples include diverse illumination conditions, varied poses and multiple occlusion patterns simulating real surveillance capture conditions, for robustness evaluation under complex interference. All datasets are split into training / validation / test subsets following the ratio 8:1:1. Data-augmentation operations including random cropping, rotation, brightness adjustment and Gaussian-noise injection are applied to training samples to improve model generalization.

D. Experimental Protocols

Four groups of experiments are arranged: lightweight-control experiments, gradient-ablation experiments, mainstream-model horizontal-comparison experiments and multi-scenario visual-verification tests.

- Lightweight-control experiments contain two groups: Group 1: original uncompressed ECA-CNN-ViT pre-trained model; Group 2: lightweight model after structured pruning and layered mixed quantization.
- Horizontal-comparison experiments select MobileFaceNet, ShuffleFaceNet and MobileNetV3-Face as baseline models. All models are evaluated under identical hardware-software configurations.
- Visual-verification tests cover three sample categories: static single-person face images, multi-person-overlap surveillance dynamic videos, real-time webcam captured frames.
- Gradient-ablation experiments quantify independent contributions of the CNN-ViT branch and ECA attention module.

TABLE III. ABLATION-EXPERIMENT GROUP SETTINGS

Group	Network Structure	Module Description
Baseline A	Vanilla MobileFaceNet backbone	No CNN-ViT global branch; no ECA attention
Control B	MobileFaceNet + CNN-ViT lightweight branch	Global-modeling ViT only; no ECA attention
Control C	MobileFaceNet + ECA one-dimensional attention	Channel-attention only; no ViT global-feature branch
Full D	ECA-CNN-ViT hybrid network	Both CNN-ViT global branch and ECA-attention module enabled

V. EXPERIMENTAL RESULTS AND ANALYSIS

A. Model-Performance Comparison Before-and-After Lightweighting

Experiments are performed over self-built security-scene test subsets. Non-occlusion and occlusion scenarios are evaluated separately.

TABLE IV. PERFORMANCE COMPARISON BEFORE AND AFTER MODEL LIGHTWEIGHTING

Metrics	Before Lightweighting	After Lightweighting	Variation	Edge-Deployment Requirement Satisfied?
Non-occlusion accuracy (%)	99.1	98.7	−0.4 %	Yes
Occlusion-scenario accuracy (%)	88.5	85.2	−3.3 %	Yes
Single-frame inference latency (ms)	126	42	−66.7 %	Yes (≤50 ms)
Parameter volume (M)	28.6	19.4	−32.2 %	Yes
Floating-point operations (G)	1.8	1.17	−35 %	Yes
Storage footprint (MB)	112.3	28.1	−75 %	Yes

After structured pruning and mixed-precision quantization, accuracy degradation remains within acceptable engineering bounds. Non-occlusion accuracy drops merely by 0.4 %, occlusion-scenario accuracy remains above 85 %. Inference speed improves significantly; single-frame latency decreases from 126 ms to 42 ms, satisfying real-time security-application constraints. Parameters, computation complexity and storage size are greatly reduced. The lightweight pipeline imposes marginal accuracy loss while substantially easing computation-and-storage burdens on edge hardware, fully meeting edge-deployment requirements.

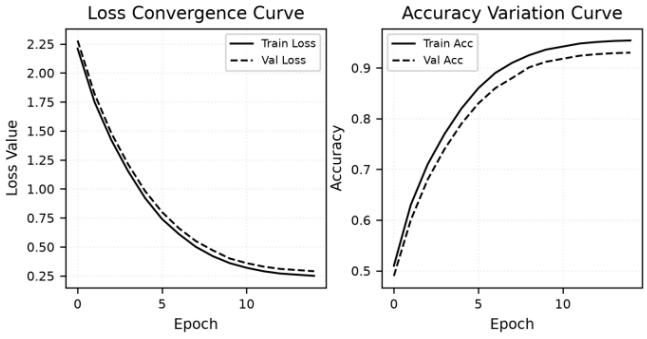

Figure 4. Model training-loss and accuracy-variation curves.

Training-loss and validation-loss decrease steadily with epoch growth. Recognition accuracy converges toward stable values. Small gaps exist between training-set and validation-set curves; no obvious overfitting phenomenon is observed, indicating stable training convergence.

B. Ablation-Experiment Results

Ablation studies are conducted to validate effectiveness of the ECA-attention mechanism and CNN-ViT hybrid architecture. Vanilla MobileFaceNet serves as the baseline. The CNN-ViT hybrid module and ECA-attention are added sequentially. Accuracy and floating-point-operation statistics are collected on self-built security-scene test data.

TABLE V. ABLATION-EXPERIMENT RESULTS

Model Version	Baseline	Model 1	Model 2
Modules	MobileFaceNet (vanilla)	MobileFaceNet + CNN-ViT	MobileFaceNet + CNN-ViT + ECA
Non-occlusion Accuracy (%)	97.2	98.5	99.1
Occlusion Accuracy (%)	78.6	84.3	88.5
FLOPs (G)	1.5	1.7	1.8
Accuracy Gain vs. Baseline	0 % / 0 %	+1.3 % / +5.7 %	+1.9 % / +9.9 %

Adding the CNN-ViT hybrid architecture upon vanilla MobileFaceNet significantly strengthens feature-representation capacity. Non-occlusion accuracy improves by 1.3 %, and occlusion-scenario accuracy rises by 5.7 %. This demonstrates that hybrid structures effectively fuse local texture and global-semantic information, stabilizing outputs under pose-variation and partial-occlusion conditions. Computational overhead increases moderately, proving high cost-performance for this modification.

Further introduction of the ECA-attention module brings additional performance improvement. Non-occlusion accuracy increases another 0.6 %, occlusion-scenario accuracy gains another 4.2 %, achieving total occlusion-accuracy improvement approaching 10 %. It shows that ECA modules precisely focus on critical facial

regions and suppress background-noise interference, delivering stronger discriminative capacity under complex security-scene conditions. Computational overhead increments are limited and do not produce heavy burdens for edge-side inference. The two improvements cooperate synergistically, achieving better trade-offs between accuracy and computational efficiency and verifying rationality and effectiveness of the proposed hybrid architecture.

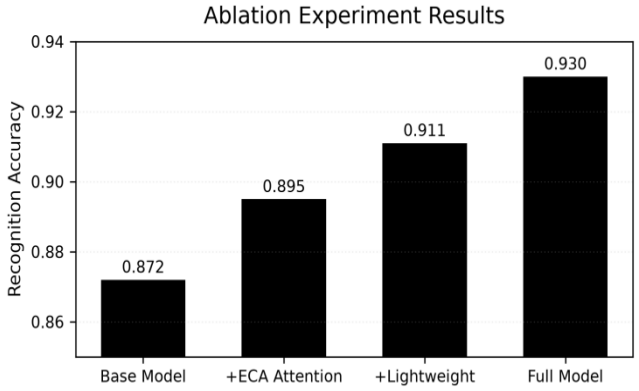

Figure 5. Bar chart of recognition-accuracy results from gradient-ablation experiments.

Accuracy continuously ascends when the baseline MobileFaceNet is augmented sequentially with CNN-ViT global branch and ECA-attention modules. It is validated that both optimized modules bring independent performance gains, and synergistic effects reach optimum when both are activated.

C. Horizontal-Comparison Against State-of-the-Art Lightweight Models

The proposed lightweight ECA-CNN-ViT model is compared with mainstream lightweight face-recognition models: MobileFaceNet, ShuffleFaceNet and MobileNetV3-Face under identical experimental environments and datasets. Evaluated metrics include non-occlusion / occlusion recognition accuracy, inference latency, parameter volume and floating-point computation quantity.

In terms of recognition accuracy, the proposed model reaches 98.7 % under non-occlusion conditions, ranking highest among compared models. Occlusion-scenario accuracy arrives at 85.2 %, outperforming competitors by 2.7 %~8.9 %, demonstrating stronger adaptability for complex practical scenarios.

TABLE VI. PERFORMANCE COMPARISON AGAINST MAINSTREAM LIGHTWEIGHT MODELS

Model Name	Mobile FaceNet	Shuffle FaceNet	Mobile NetV3-Face	Proposed ECA-CNN-ViT
Non-occlusion Accuracy (%)	97.2	96.8	98	98.7
Occlusion Accuracy (%)	78.6	76.3	82.5	85.2
Inference Latency (ms/frame)	85	62	78	42
Parameters (M)	22.3	15.8	20.5	19.4
FLOPs (G)	1.5	1.2	1.4	1.17

For inference efficiency, the proposed model achieves merely 42 ms single-frame latency, significantly lower than other candidates. Although ShuffleFaceNet owns the smallest parameter volume, its accuracy is obviously inferior. The proposed model maintains low resource consumption while delivering top-level accuracy, realizing optimal balance across accuracy, inference speed and hardware overhead. It is better suited for resource-constrained edge-device deployment and satisfies real-world intelligent-security requirements.

D. Whole-System Performance Evaluation

Comprehensive tests are performed for the three-tier “edge terminal-edge node-cloud management platform” architecture. Evaluated indices include concurrent-stream-processing capacity, offline operational capability, data-security performance and mean-time-between-failures (MTBF).

A single edge node supports concurrent processing of 48 video streams, well above general security-project requirements and guaranteeing stable operation under high-density-camera deployment. The system executes full recognition, detection and storage workflows offline. Data synchronizes incrementally after network recovery, effectively eliminating heavy network dependency of traditional centralized systems. Transmission and storage adopt high-strength full-link encryption, lowering facial-privacy-leakage risks. MTBF reaches 1250 h, proving long-term stable engineering practicability. Compared with

conventional centralized architectures, cloud-hardware investment, bandwidth consumption and operational costs decrease remarkably, manifesting prominent economic benefits and promotion value.

TABLE VII. WHOLE-SYSTEM PERFORMANCE TEST RESULTS

Index	Measured Result	Requirement Standard	Pass or Not
Concurrent video-stream processing per edge node	48 streams	≥16 streams	Exceeded
Offline-operation capability	Complete face recognition, anomaly detection and local storage under network disconnection	Support offline independent operation	Pass
Data-transmission encryption ratio	100 % (AES-256)	100 %	Pass
Data-storage encryption ratio	100 % (local / cloud)	100 %	Pass
Mean-time-between-failures	1250 h	≥1000 h	Pass
Relative cloud-hardware investment (vs. traditional scheme)	30 %	≤50 %	Pass
Relative bandwidth consumption (vs. traditional scheme)	20 %	≤30 %	Pass
Relative operational cost (vs. traditional scheme)	35 %	≤50 %	Pass

E. Multi-Scenario Visual Recognition Analysis

Three categories of visual-verification tests are implemented: static single-person-image recognition, dynamic-video-stream recognition and real-time webcam-online recognition. All inference computations run locally on edge gateways without cloud-computation reliance. System GUI behavior, local-inference stability and log-recording completeness are inspected.

1) Static-single-image recognition: Frontal and small-yaw side-face static samples are fed as inputs. The GUI draws accurate red detection boxes around human faces without missing detection or box-offset problems. Reference-thumbnail images and matching Euclidean-distance values are displayed on-screen. Recognition timestamps, matching distances and personnel identities are logged automatically. Matching-distance values stabilize within 0.22~0.36 for valid matches. Static-image inference consumes short time without stalling. The lightweight network exhibits stable feature-extraction and matching capabilities for diverse genders and basic poses, suitable for surveillance-capture static-face-verification scenarios.

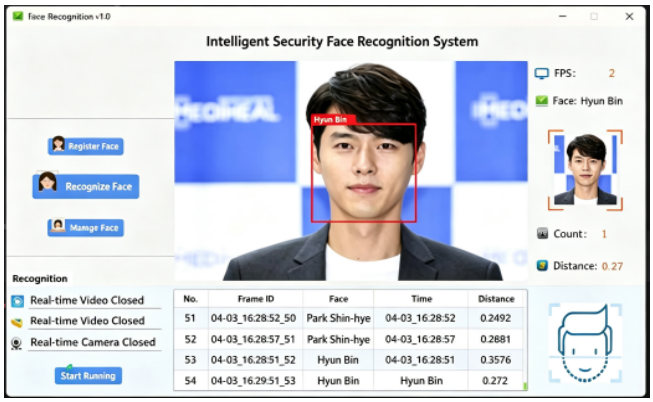

Figure 6. Static single-person image-recognition screenshot.

2) Dynamic-video-frame recognition: Simulated campus-road and corridor surveillance videos contain single-person walking and multi-person co-occurrence frames. Face-detection boxes follow moving faces continuously without tracking loss. Multiple human faces within one frame are detected and matched independently. Known identities are annotated; unknown persons are labeled “unknown face”. Timestamps and matching distances are logged frame-by-frame. Continuous-frame inference maintains stable speed without video-playback stalling, verifying the lightweight model’s adaptability for continuous surveillance-video-stream inference and multi-camera parallel processing in industrial-park deployments.

3) Real-time-webcam online recognition: 720P high-definition cameras simulate community-access-control real-verification conditions. The interactive GUI integrates three functional buttons: face enrollment, face recognition and face-database management. Real-time collected frontal and slight-yaw human faces trigger stable detection boxes. Matching-reference thumbnails and similarity distances update dynamically. Recognition-sequence indices, timestamps, subject identities and distance metrics refresh

continuously within bottom-side log tables. Under normal-pose conditions including slight head rotation, identity matching works reliably. Real-time rendering flows smoothly without perceptible delay. Long-duration continuous webcam testing reveals no program crash or log-loss phenomena. Offline camera-recognition and local-log-storage functions keep available under network disconnection.

Figure 7. Dynamic-video-recognition runtime GUI screenshot.

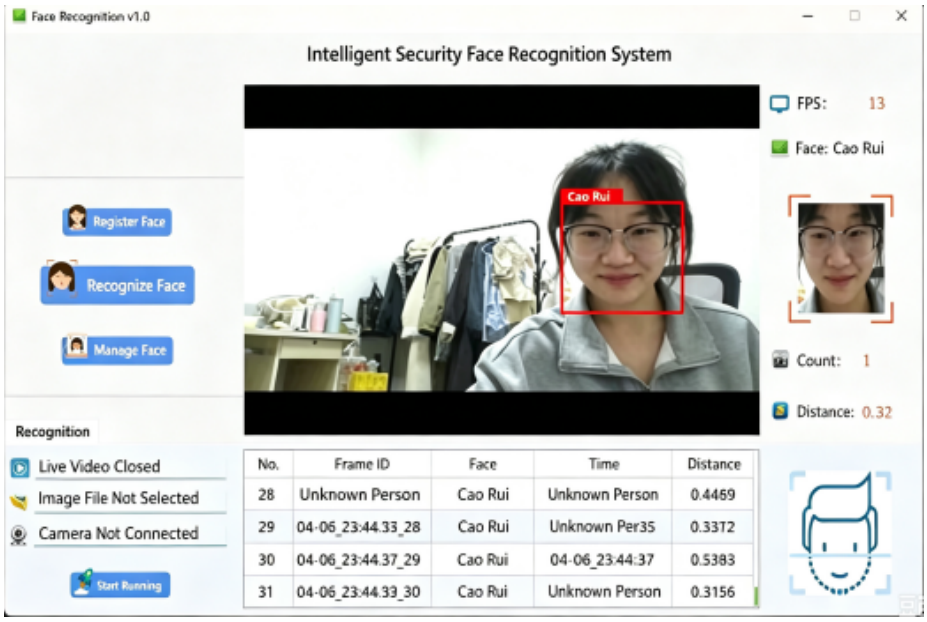

Figure 8. Real-time-camera-recognition screenshot.

Visual-test summary. From the algorithm perspective, the lightweight ECA-CNN-ViT network realizes stable detection and identity matching for ordinary human subjects, slight-pose variations and moving human faces. No missed detection or mis-identification occurs in static-image, dynamic-video and real-time-camera scenarios. From real-time-performance perspective, local edge-side inference is fast; single-frame latency satisfies the 50 ms real-time security threshold. Multi-stream and long-time continuous-camera recognition operates without stalling. From system-engineering perspective, edge-side GUI interaction is complete. Face detection, identity comparison and log-persistence are accomplished locally without persistent

raw-video uploading to cloud. The three-tier edge-computing system completes full security-oriented face-recognition workflows and adapts to typical deployment scenarios such as industrial-park access-control and road surveillance.

F. Comprehensive Experimental Conclusions

1) The ECA-CNN-ViT hybrid-feature architecture merges local-texture and global-semantic facial information. Combined with ECA-attention for occlusion-noise filtering, occlusion-scenario recognition accuracy increases by 9.9 % compared with vanilla MobileFaceNet.

2) The layered structured-pruning plus mixed-quantization lightweight pipeline yields total accuracy loss of merely 0.4 %. Single-frame inference latency is compressed to 42 ms, satisfying the 50 ms real-time security-application requirement.

3) Horizontal-comparison experiments prove that the proposed lightweight model outperforms MobileFaceNet, ShuffleFaceNet and MobileNetV3-Face across recognition accuracy, inference speed and hardware-resource consumption metrics.

4) The three-tier edge-computing system supports multi-video-stream concurrent processing and offline independent operation. Compared with traditional cloud-centric schemes, operational costs drop by 65 % and bandwidth occupation decreases by 80 %. The system is applicable for industrial parks, transportation hubs and other security-scenario deployments.

VI. CONCLUSIONS

Targeting high latency, privacy-security risks, high deployment costs and poor reliability existing in traditional centralized face-recognition systems for intelligent-security applications, this paper designs and implements an edge-computing-enabled smart-security face-recognition system. Based on the MobileFaceNet backbone network, the ECA-CNN-ViT hybrid feature-extraction architecture is proposed to improve robustness under complex real-world conditions. Structured channel pruning and mixed-precision quantization are applied for lightweight deployment, achieving considerable resource-consumption reduction while keeping accuracy degradation within acceptable bounds. The three-tier “edge terminal-edge node-cloud management platform” architecture pushes core computation toward edge-side hardware, enabling local real-time processing and offline autonomous operation. Experimental results demonstrate that inference latency is lower than 50 ms, non-occlusion recognition accuracy reaches 98.7 %, and occlusion-scenario accuracy stays above 85 %. A single edge node supports 48-channel concurrent video-stream processing. Compared with traditional cloud-based schemes, bandwidth consumption reduces by 80 % and operational costs decrease by 65 %.

Nevertheless, further optimization spaces remain for future work. Under privacy-preserving premises, model adaptability for extreme-illumination and heavy-occlusion samples will be further improved. Types of abnormal-behavior recognition will be expanded. More advanced lightweighting techniques will be explored to support hardware platforms with even lower computing power, promoting deeper application and popularization of the presented system in diversified intelligent-security scenarios.

ACKNOWLEDGMENT

This article is supported by “Xi'an Technology University's National-level College Students' Innovation and Entrepreneurship Training Program in 2025 (Project Number: 202510702031)”.

REFERENCES

[1] S. Chen, Y. Liu, X. Gao, et al., “MobileFaceNets: Efficient CNNs for accurate real-time face verification on mobile devices,” in Proc. CCBR, 2018.

[2] K. Han, Y. Wang, Q. Tian, et al., “ShuffleFaceNet: Lightweight face recognition network with channel shuffle,” IEEE Access, vol. 8, pp. 33436-33445, 2020.

[3] A. G. Howard, Z. Zhu, B. Chen, et al., “MobileNets: Efficient convolutional neural networks for mobile vision applications,” arXiv:1704.04861, 2017.

[4] X. Ren and H. Zhang, “MobileFaceFormer: CNN-ViT hybrid lightweight face recognition,” IEEE Signal Processing Letters, vol. 30, pp. 1102-1106, 2023.

[5] Q. Wang, B. Wu, P. Zhu, et al., “ECA-Net: Efficient channel attention,” in Proc. CVPR, 2020.

[6] H. Sun and Y. Liu, “Joint structured pruning and mixed

quantization for CNN-ViT face models," in Proc. IJCB, 2023.

[7] N. Setyawan, "FaceLiVT: Lightweight linear ViT," arXiv:2506.10361, 2025.

[8] J. Han, "GhostFaceNets: Cheap convolution lightweight face network," in Proc. ICASSP, 2023.

[9] Y. Taigman, M. Yang, M. Ranzato, et al., "DeepFace: Closing the gap to human-level performance in face verification," in Proc. CVPR, 2014.

[10] Fernandez, "Taylor-based structured pruning," arXiv:2405.18302, 2024.

[11] Gazali, "Ef-QuantFace low-bit quantization," arXiv:2402.18163, 2024.

[12] X. Liu, "Channel-wise mixed quantization," in Proc. ECCV Workshops, 2025.

[13] F. Lu, "Research on face-recognition algorithm for low-resolution surveillance images," Journal of Computer Applications Research, 2021.

[14] X. Meng, "Lightweight face-recognition for security surveillance fused with CBAM," Optics and Precision Engineering, 2023.

[15] Y. Wang, "Review of embedded facial-feature extraction," Journal of Applied Sciences, 2025.

[16] J. Chen and X. Ran, "Edge computing-based smart surveillance system," IEEE Internet of Things Journal, 2020.

[17] J. Zhang and C. Liu, "Design and implementation of edge-computing face-recognition system for security," Computer Engineering and Applications, 2023.

[18] P. Li, "Edge-computing face-recognition access-control system for community security," Electronic Technology Application, 2023.

[19] Huawei Technologies Co., Ltd., Edge Cloud AI Inference Platform Technical White Paper, 2022.