BFS-CNN-ECA-GMP-GRU-MSP: An Enhanced Cross-Perspective Gait Recognition Model with Efficient Channel Attention and Cosine-Consistent Metric Learning

Lijuan Gao[1,2*] and Jieran Liu[1]
[1]School of Artificial Intelligence, Zhengzhou University of Industrial Technology, Zhengzhou, 451150, China
[2]Henan Engineering Technology Research Center of Intelligent Transportation Video Image Perception and Recognition, Zhengzhou, 451150, China
E-mail: lijuan_gao220@163.com, jieran_liu@163.com
*Corresponding author

Keywords: Cross-perspective, gait recognition, efficient channel attention, BFS algorithm, cosine metric learning, feature map interaction

Received: March 3, 2026

Cross-view gait recognition remains vulnerable to viewpoint shifts and appearance changes, especially under carrying and clothing covariates. We propose BFS-CNN-ECA-GMP-GRU-MSP, an enhanced version of our previous BFS-CNN-GMP-GRU-MSP framework, by introducing two upgrades: multi-stage lightweight channel recalibration with Efficient Channel Attention (ECA) and cosine-consistent metric learning through cosine batch-hard triplet loss, a cosine classifier, and L2-normalized embeddings. Experiments are first conducted on CASIA-B under the same legacy closed-set protocol used in our earlier Informatica study (gallery: NM#01–04; probe: NM#05–06, BG#01–02, CL#01–02; same-view matches excluded). This protocol is retained to isolate architecture-level improvements and is interpreted as a within-protocol comparison rather than a subject-disjoint generalization benchmark. Under this setting, the proposed model reaches mean Rank-1 accuracies of 99.96% (NM), 99.74% (BG), and 98.38% (CL), improving the baseline by +2.96, +5.74, and +7.38 percentage points, respectively. To probe unseen-subject behavior more directly, we further report a supplementary subject-disjoint split (subjects 001–074 for training and 075–124 for testing), where the full model attains 97.73% (NM), 87.98% (BG), and 64.00% (CL). Under this stricter split, the clearest effect of ECA appears under clothing variation, where the full model exceeds w/o ECA by 1.82 percentage points on CL, while the ECA branch still introduces only 13 learnable parameters (k=3/5/5 for 64/128/256 channels). These results support the proposed modifications as a lightweight and effective enhancement for protocol-matched cross-view gait recognition, while broader multi-split subject-disjoint and open-set validation remains future work.

Povzetek: Članek predstavi lažjo nadgradnjo CNN–GRU okvira za prepoznavo hoje med pogledi, ki z ECA kanalno pozornostjo in kosinusno konsistentnim metričnim učenjem izboljša robustnost pri nošenju predmetov in spremembah oblačil, ob zelo majhnem številu dodatnih parametrov.

1 Introduction

Gait recognition is a biometric technology that identifies individuals based on their walking patterns. Compared with other biometric modalities such as face, fingerprint, and iris recognition, gait recognition can operate at a distance without requiring subject cooperation and does not rely on high-resolution image acquisition devices [2, 4]. These properties make gait recognition particularly valuable for applications in video surveillance, public security monitoring, and forensic investigation. In recent years, deep learning methods have advanced gait recognition substantially, with convolutional and recurrent networks being widely adopted for spatial and temporal feature extraction from gait sequences [3, 4].

Despite this progress, cross-perspective gait recognition remains difficult in real deployments. The same subject can look very different across camera viewpoints, and covariates such as bags and heavy coats further distort body contours [5, 6]. In CASIA-B, the clothing condition (CL) is typically the hardest setting because the coat changes upper-body shape and suppresses discriminative silhouette cues [7].

Existing methods address this issue from different angles. GaitSet [8] models a sequence as an unordered set, GaitPart [9] learns part-aware temporal dynamics, and GaitGL [10] combines global and local representations. Recent work such as DyGait [11] and QAGait [12] further improves dynamic and quality-aware modeling. OpenGait [3] has also made protocol and implementation comparison more systematic. Even with these advances, robustness under severe covariates and cross-view shifts is still a practical

bottleneck.

In our previous work [1], we introduced a BFS-CNN-GMP-GRU-MSP framework that combines multi-scale spatial extraction, BFS-style neighborhood propagation, and GRU-based temporal modeling. That model reached 0.97/0.94/0.91 accuracy under NM/BG/CL, but two limitations remained. First, channel recalibration was missing, so the model could not explicitly suppress noisy channels under strong covariates. Second, the training objective and evaluation metric were inconsistent: triplet loss used Euclidean distance, while retrieval used cosine similarity.

To address these limitations, we propose an enhanced model, BFS-CNN-ECA-GMP-GRU-MSP, with three contributions:

(1) Lightweight channel recalibration with ECA. Efficient Channel Attention (ECA) modules [13] are inserted after each CNN stage. Compared with SE-style channel attention [14], ECA uses local channel interaction through a 1D convolution and introduces only 13 parameters across all three stages.

(2) End-to-end cosine-consistent metric learning. We align training and inference in the same angular space: cosine-distance batch-hard triplet loss, cosine classifier, and L2-normalized embeddings.

(3) Controlled empirical validation. Under the same CASIA-B protocol used by our baseline [1], we report full-condition results, per-view analysis, ablation studies, and feature visualizations (t-SNE and Grad-CAM), and we additionally provide a supplementary subject-disjoint split to examine unseen-subject behavior more directly.

Research objective and scope. The goal of this study is not to claim universal state-of-the-art performance across all gait benchmarks, but to answer a narrower and practically relevant question: *can lightweight channel attention and cosine-consistent metric learning improve cross-view gait recognition under appearance covariates while keeping additional parameter cost negligible?* Our working hypothesis is that explicitly reweighting channels and aligning the optimization metric with the retrieval metric will be especially beneficial under the clothing condition (CL), where silhouette distortion is strongest.

The remainder of this paper is organized as follows. Section 2 reviews related work on gait recognition, attention mechanisms, and metric learning. Section 3 presents the proposed method in detail. Section 4 reports the experimental results and analysis. Section 5 discusses the broader implications, limitations, and practical considerations of the study. Section 6 concludes the paper.

2 Related works

2.1 Cross-view gait recognition

Cross-view gait recognition has attracted considerable research attention over the past decade. Early methods primarily relied on handcrafted features and geometric transformations to handle viewpoint changes, but these approaches generally suffered from limited representational capacity [2]. With the rise of deep learning, CNN-based methods have become the dominant paradigm. Chao et al. [8] proposed GaitSet, which regards the gait sequence as an unordered set and uses set pooling to aggregate frame-level features. This approach is robust to sequence length variations but ignores the temporal ordering information that is inherent in walking patterns.

Fan et al. [9] introduced GaitPart, a temporal part-based model that divides the silhouette into horizontal strips corresponding to different body parts and applies separate temporal modeling to each part. This design captures the distinct motion patterns of different body regions, such as arm swing and leg stride. Lin et al. [10] proposed GaitGL, which integrates global and local feature representations through a unified framework and achieves local temporal aggregation to capture short-range temporal dynamics. More recently, DyGait [11] explicitly targets the extraction of dynamic features from moving body parts, achieving strong results on multiple benchmark datasets. QAGait [12] addresses gait recognition from a quality-aware perspective, handling low-quality samples that often degrade performance in practical scenarios. Fan et al. [16] proposed SkeletonGait, which generates skeleton-based gait maps as an alternative modality to silhouettes, providing robustness against appearance-level noise.

Our previous work [1] proposed the BFS-CNN-GMP-GRU-MSP model that introduces breadth-first search for feature propagation across the feature map, combined with multi-scale spatial pyramid fusion and GRU-based temporal attention. While this model demonstrated the effectiveness of BFS-based feature interaction, it did not incorporate channel-level attention and used inconsistent distance metrics between training and evaluation.

Table 1 provides a contextual reference for representative CASIA-B methods. Two observations motivate the present work. First, the literature shows that performance gains often arrive together with broader backbone redesign, protocol tuning, or auxiliary quality processing. Second, relatively less attention has been paid to whether the training objective and the final retrieval metric are geometrically aligned. Our work therefore targets a narrower gap: improving a protocol-matched baseline through lightweight channel recalibration and cosine-consistent optimization without materially increasing model size.

2.2 Attention mechanisms

Attention mechanisms have been widely applied in computer vision to enhance feature representations by selectively focusing on informative regions or channels [17]. The Squeeze-and-Excitation Network (SE-Net) [14] introduced the concept of channel attention by using global average pooling followed by two fully connected layers to generate channel-wise weights. While effective, the SE block introduces a non-negligible number of parameters due to its dimensionality reduction design.

Table 1: Protocol-aware contextual comparison of representative CASIA-B gait methods. Reported numbers are taken from the original papers under their own settings and are therefore shown for contextual reference only rather than strict apples-to-apples comparison. Because parameter counts and runtime settings are not reported consistently across the cited sources, they are discussed separately rather than forced into the same table. "N/R" denotes not reported in the cited source.

Method	Venue	Protocol / setting note	NM	BG	CL	Mean
GaitSet [8]	AAAI'19	Original CASIA-B setting reported in the source paper	95.0	87.2	70.4	84.2
GaitPart [9]	CVPR'20	Identical-view cases excluded in the original paper	96.2	91.5	78.7	88.8
GaitGL [10]	ICCV'21	Global-local plus local temporal aggregation (LT setting)	97.4	94.5	83.6	91.8
DyGait [11]	ICCV'23	Mean CASIA-B Rank-1 reported in the abstract; per-condition numbers not separately listed in the source paper	N/R	N/R	N/R	98.4
QAGait [12]	AAAI'24	Quality-aware setting on a reduced GaitBase backbone for small-scale CASIA-B	97.9	94.6	78.2	90.2
Previous baseline [1]	Informatica 2022	Legacy protocol retained in the present study for direct comparison	97.0	94.0	91.0	94.0

Wang et al. [13] proposed ECA-Net (Efficient Channel Attention Network), which replaces the fully connected layers in SE-Net with a single one-dimensional convolution. The kernel size of the convolution is adaptively determined based on the number of channels, enabling local cross-channel interaction without dimensionality reduction. This design achieves comparable or superior performance to SE-Net while significantly reducing the parameter overhead. The simplicity and efficiency of ECA make it particularly suitable for integration into existing architectures where parameter budget is a concern.

In the context of gait recognition, attention mechanisms have been explored for both spatial and temporal feature refinement. Yan et al. [7] proposed adaptive structured spatial representations combined with multi-scale temporal aggregation. However, the systematic application of lightweight channel attention specifically to multi-stage gait feature extraction has not been fully explored.

2.3 Metric learning for person identification

Metric learning plays a crucial role in person re-identification and gait recognition tasks. The triplet loss, originally popularized by Hermans et al. [18] through the batch-hard mining strategy, encourages the model to learn embeddings where samples of the same identity are closer together than samples of different identities. The choice of distance metric in the triplet loss significantly affects training dynamics and convergence.

Wojke and Bewley [19] demonstrated the advantages of cosine distance over Euclidean distance for person re-identification, showing that cosine metric learning produces more discriminative embeddings in the angular space. Luo et al. [20] proposed the BNNeck design and a bag of tricks for person re-identification, including batch normalization before the classification head and separate treatment of features for triplet and cross-entropy losses. Their work established a strong baseline that has been widely adopted in subsequent research.

Label smoothing [21], originally proposed for training deep classification networks, has been shown to prevent overconfident predictions and improve generalization. In the context of metric learning, label smoothing applied to the cross-entropy branch helps regularize the training process and produces more calibrated classification scores.

3 Proposed method

This section presents the proposed BFS-CNN-ECA-GMP-GRU-MSP model in detail. The overall architecture is first described, followed by the design of each component module.

3.1 Difference from the previous Informatica model

The present manuscript extends rather than replaces our earlier BFS-CNN-GMP-GRU-MSP framework [1]. The inherited components are the three-stage CNN backbone, BFS-based feature propagation, MSP spatial fusion, GRU-GMP temporal modeling, and the gated fusion design. The new elements introduced in this paper are: (1) multi-stage ECA channel attention after each backbone stage; and (2) a cosine-consistent learning pipeline that replaces the Euclidean triplet objective and standard linear classification

geometry with cosine-distance batch-hard triplet loss, cosine classification, and L2-normalized embeddings. This explicit separation clarifies that the current contribution lies in a lightweight architectural enhancement and a training/evaluation metric redesign built on a previously established backbone.

3.2 Overall architecture

The proposed model takes gait silhouette sequences as input and produces identity embeddings for recognition. The input is a batch of silhouette sequences with dimensions $[B, T, 1, H, W]$, where B is the batch size, T is the number of frames (set to 30), $H = 64$ and $W = 44$ are the spatial dimensions. The architecture consists of seven main components arranged in a sequential pipeline:

(1) CNN Backbone: A three-stage residual convolutional network that extracts hierarchical spatial features at multiple scales, producing feature maps of 64, 128, and 256 channels respectively.

(2) ECA Channel Attention: Efficient Channel Attention modules applied after each CNN stage to dynamically recalibrate channel-wise feature responses.

(3) BFS Feature Propagation: Breadth-first search inspired neighborhood aggregation modules that propagate information across the spatial extent of each feature map with gated fusion.

(4) Multi-Scale Spatial Pyramid (MSP): A multi-scale dilated convolution module that fuses the three-level features through horizontal part-based pooling to produce spatial feature vectors for each body part.

(5) GRU-GMP Temporal Attention: A temporal modeling module that combines global max pooling, bidirectional GRU, and learnable temporal attention to capture gait dynamics.

(6) Gated Fusion: A learnable gating mechanism that adaptively combines the spatial and temporal feature representations.

(7) BNNeck and Cosine Classifier: Batch normalization neck followed by cosine-based classification, producing both class predictions and L2-normalized embeddings.

The model outputs three tensors: triplet features (before batch normalization, used for triplet loss computation), classification scores (from the cosine classifier), and L2-normalized embedding vectors (for evaluation retrieval). The total number of trainable parameters is 6.83M.

3.3 CNN backbone

The CNN backbone consists of three stages, each containing a basic convolution block followed by a residual stage. The basic convolution block applies a 3×3 convolution, batch normalization, and LeakyReLU activation (negative slope 0.2). The residual stage consists of two 3×3 convolutional layers with batch normalization and LeakyReLU, employing a skip connection. When the input and output channel dimensions differ, a 1×1 convolution is used to align the residual branch.

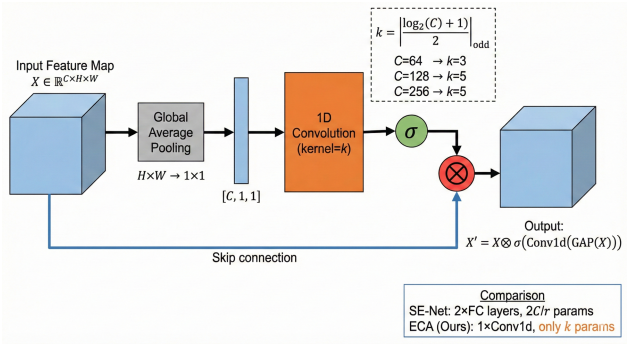

Figure 1: Overall architecture of the proposed BFS-CNN-ECA-GMP-GRU-MSP model

The three stages operate as follows:

- **Stage 1**: Input channel 1 → output channel 64, spatial dimensions preserved, followed by 2×2 max pooling.
- **Stage 2**: Input channel 64 → output channel 128, followed by 2×2 max pooling.
- **Stage 3**: Input channel 128 → output channel 256, no additional downsampling.

The output feature maps of the three stages are $[B \times T, 64, H, W]$, $[B \times T, 128, H/2, W/2]$, and $[B \times T, 256, H/4, W/4]$ respectively, where each stage captures spatial information at a different level of abstraction: shallow texture features, mid-level structural features, and deep semantic features.

3.4 Efficient channel attention (ECA)

The ECA module is the first key innovation of this work. It is applied independently after each stage of the CNN backbone to perform channel-wise feature recalibration. The computation proceeds as follows.

Given an input feature map $X \in \mathbb{R}^{C\times H\times W}$, global average pooling is first applied to aggregate spatial information:

$$y = \mathrm{GAP}(X) \in \mathbb{R}^{C\times 1\times 1}$$

Then, a one-dimensional convolution with kernel size k is applied along the channel dimension to capture local cross-channel interactions:

$$\hat{y} = \sigma(\mathrm{Conv1d}_k(y))$$

where σ denotes the sigmoid activation function. The output is used as channel-wise attention weights to recalibrate the input feature map:

$$\mathrm{ECA}(X) = X \odot \hat{y}$$

where $\odot$ denotes element-wise multiplication with broadcasting over the spatial dimensions.

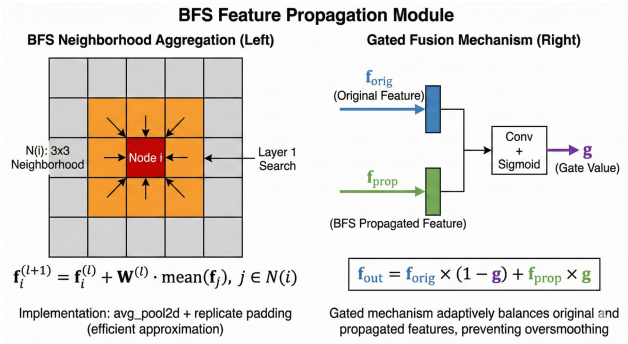

Figure 2: Structure of the efficient channel attention (ECA) module

The kernel size k is adaptively determined based on the number of channels C using the following formula:

$$k = \left| \frac{\log_2 C + b}{\gamma} \right|_{\text{odd}}$$

where $\gamma = 2$ and $b = 1$. The subscript "odd" indicates that the result is rounded to the nearest odd integer.

The parameter overhead introduced by ECA is remarkably small. Table 1 summarizes the configuration at each stage:

The ECA modules add 13 learnable parameters in total (0.0002% of 6.83M). This keeps the attention branch lightweight while retaining measurable gains under challenging covariates. The structure is shown in Figure 2.

3.5 BFS feature propagation

The BFS feature propagation module, originally introduced in our previous work [1], enables global information exchange across the feature map by simulating breadth-first search through neighboring nodes. Each spatial position in the feature map is treated as a node, and the BFS process propagates information from each node to its neighbors in a layer-by-layer manner.

For a single propagation layer, the feature update rule is:

$$f_i^{(l+1)} = f_i^{(l)} + \mathbf{W}^{(l)} \cdot \frac{1}{|\mathcal{N}(i)|} \sum_{j \in \mathcal{N}(i)} f_j^{(l)} \qquad (1)$$

where $f_i^{(l)}$ is the feature of node i at search layer l, $\mathcal{N}(i)$ is the 3×3 spatial neighborhood of node i, and $\mathbf{W}^{(l)}$ is a learnable weight matrix implemented as a convolution with channel reduction ratio of 4.

A gated fusion mechanism is employed to balance the original features and the propagated features:

$$g = \sigma\left(\text{Conv}([f_{\text{orig}}, f_{\text{prop}}])\right)$$

$$f_{\text{out}} = f_{\text{orig}} \cdot (1 - g) + f_{\text{prop}} \cdot g$$

where $[\cdot, \cdot]$ denotes channel-wise concatenation and g is the gating value computed through a convolutional layer followed by sigmoid activation.

3.6 Multi-scale spatial pyramid (MSP)

The MSP module fuses the three-level feature maps into a unified spatial feature representation through multi-scale dilated convolutions and horizontal part-based pooling.

FrameMax temporal aggregation. For each of the three feature map levels, the temporal dimension is first collapsed by taking the element-wise maximum across all frames:

$$F_{\max} = \max_{t=1}^{T} F_t$$

Multi-scale dilated convolution. Each level of feature maps is processed by a group of dilated convolutions with dilation rates of [1, 2, 4], capturing spatial relationships at different ranges: local, medium-distance, and long-distance, respectively. The dilated convolution is an effective approach for expanding the receptive field without increasing the number of parameters [22].

Channel alignment. All three levels of features are projected to a common channel dimension of 128 through 1×1 convolutions, enabling subsequent feature fusion.

Horizontal part-based pooling (HPP). The aligned feature maps are spatially divided into $P = 8$ horizontal strips through adaptive average pooling, where each strip corresponds to a different body region (head, upper torso, lower torso, legs, etc.).

Feature concatenation and mapping. For each body part, the features from the three levels are concatenated along the channel dimension ($128 \times 3 = 384$ dimensions) and mapped to the target feature dimension (256) through a part-specific fully connected layer.

The output of the MSP module is a spatial feature tensor of size $[B, 8, 256]$, representing 8 body part features of 256 dimensions each.

3.7 GRU-GMP temporal attention module

The temporal attention module captures the temporal dynamics of gait sequences through a combination of global max pooling, bidirectional GRU, and learnable attention weights.

Global max pooling (GMP). The Stage 3 feature map is first processed by a 3×3 convolution for spatial feature refinement. Then, global max pooling is applied over the spatial dimensions (H, W) to extract the most salient channel response at each time step:

$$v_t = \max_{h,w} F_t(c, h, w)$$

Bidirectional GRU. The temporal feature sequence is fed into a bidirectional GRU [23] with hidden size 128 in each direction, producing hidden states of dimension 256:

$$\overrightarrow{h_t} = \text{GRU}_{\text{fw}}(v_t, \overrightarrow{h_{t-1}}), \quad \overleftarrow{h_t} = \text{GRU}_{\text{bw}}(v_t, \overleftarrow{h_{t+1}})$$

$$h_t = [\overrightarrow{h_t}, \overleftarrow{h_t}]$$

Table 2: ECA module configuration at each feature extraction stage

Stage	Channels C	Kernel size k	Parameters
Stage 1 (Shallow)	64	3	3
Stage 2 (Middle)	128	5	5
Stage 3 (Deep)	256	5	5
Total	—	—	**13**

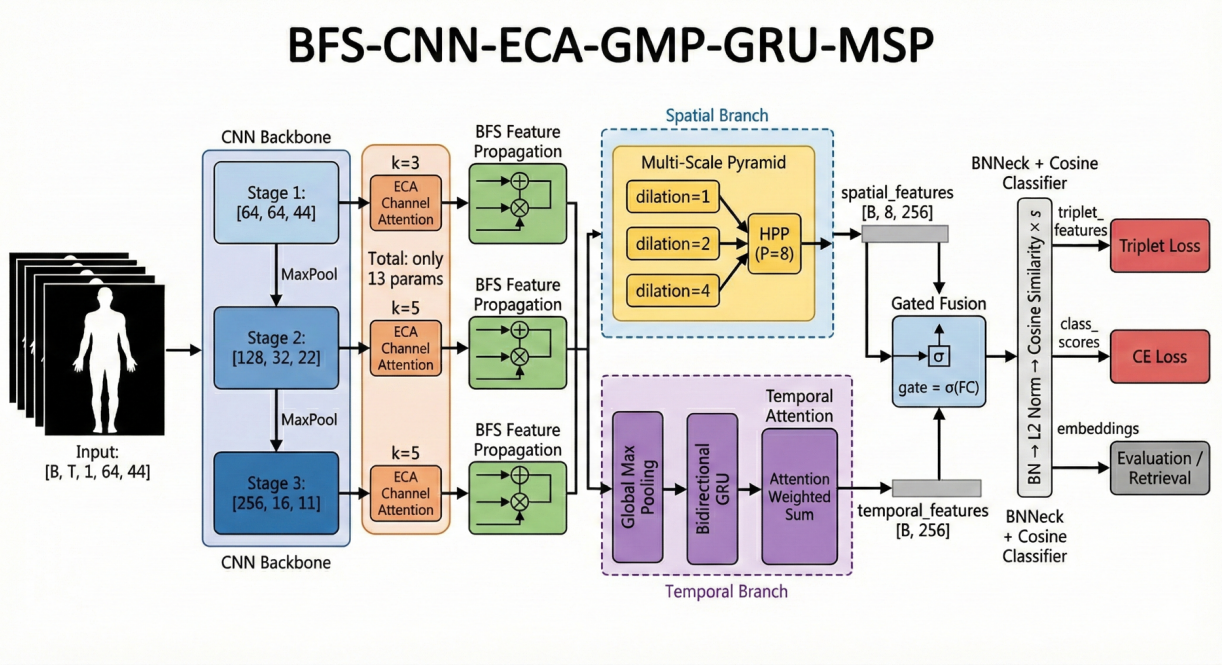

Figure 3: BFS feature propagation mechanism on a feature map

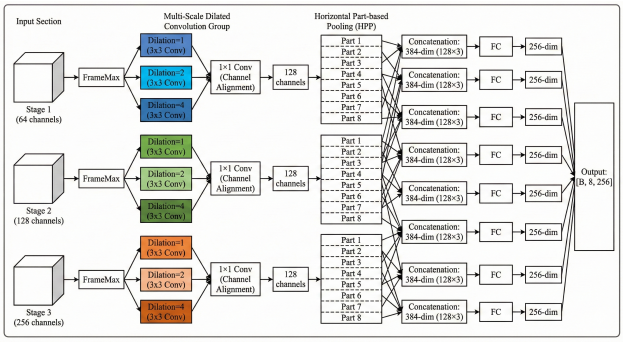

Figure 4: Structure of the multi-scale spatial pyramid (MSP) module

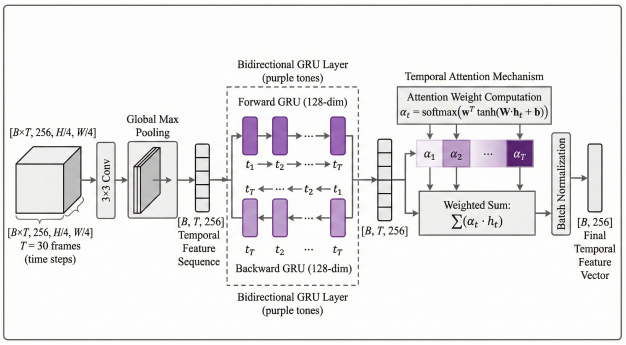

Figure 5: Structure of the GRU-GMP temporal attention module

Temporal attention. A learnable attention mechanism computes importance weights for each time step:

$$\alpha_t = \text{softmax}\left(\mathbf{w}^T \tanh(\mathbf{W}h_t + \mathbf{b})\right)$$

$$c = \sum_{t=1}^{T} \alpha_t \cdot h_t$$

3.8 Gated fusion

The spatial features from MSP (part-based, $[B, 8, 256]$) and temporal features from GRU-GMP (global, $[B, 256]$) are fused with a learnable gate:

$$g^{(p)} = \sigma\left(\mathbf{W}_g[f^{(p)}_{\text{spatial}}, f_{\text{temporal}}] + \mathbf{b}_g\right), \quad (2)$$

$$f^{(p)}_{\text{fused}} = g^{(p)} \odot f^{(p)}_{\text{spatial}} + \left(1 - g^{(p)}\right) \odot f_{\text{temporal}}, \quad (3)$$

where $p \in \{1, \ldots, P\}$.

3.9 BNNeck and cosine classifier

Following the design principle of BNNeck [20], batch normalization is applied to the fused features before the classification head. The classification head employs a cosine classifier:

$$\text{score}_{p,c} = s \cdot \frac{f_{\text{bn}}^{(p)}}{||f_{\text{bn}}^{(p)}||_2} \cdot \frac{W_c}{||W_c||_2}$$

The final embedding vector for evaluation is obtained by averaging the BN features across all 8 parts and applying L2 normalization:

$$e = \frac{\frac{1}{P}\sum_{p=1}^{P} f_{\text{bn}}^{(p)}}{\left|\left|\frac{1}{P}\sum_{p=1}^{P} f_{\text{bn}}^{(p)}\right|\right|_2}$$

3.10 Loss functions

The model is trained with a joint loss function combining two complementary objectives:

$$\mathcal{L} = \lambda_{\text{tri}} \cdot \mathcal{L}_{\text{triplet}} + \lambda_{\text{CE}} \cdot \mathcal{L}_{\text{CE}}$$

Cosine-based batch-hard triplet loss. The triplet branch uses cosine distance:

$$\mathcal{L}_{\text{triplet}} = \frac{1}{N}\sum_{i=1}^{N}[d_i^+ - d_i^- + m]_+, \tag{4}$$

where

$$d_i^+ = \max_{p \in \mathcal{P}(i)} d_{\cos}(a_i, p), \tag{5}$$

$$d_i^- = \min_{n \in \mathcal{N}(i)} d_{\cos}(a_i, n), \tag{6}$$

and $d_{\cos}(x, y) = 1 - \frac{x \cdot y}{||x||_2 \cdot ||y||_2}$. The margin is set to $m = 0.2$. The batch-hard mining strategy [18] selects the hardest positive and hardest negative within each mini-batch.

Label smoothing cross-entropy loss. The classification loss uses label smoothing [21]:

$$\mathcal{L}_{\text{CE}} = -\sum_{c=1}^{C} q_c \log p_c$$

with smoothing factor $\epsilon = 0.1$ and $C = 124$ classes. The loss weights are $\lambda_{\text{tri}} = 1.0$ and $\lambda_{\text{CE}} = 0.5$.

4 Experiments

4.1 Dataset and evaluation protocol

Dataset. All experiments are conducted on the CASIA-B gait dataset [15], which is one of the most widely used benchmarks for cross-view gait recognition. CASIA-B contains gait data from 124 subjects, each recorded from 11 viewpoints ranging from 0° to 180° with 18° intervals. Three walking conditions are included: normal walking (NM, 6 sequences per subject), carrying a bag (BG, 2 sequences), and wearing a coat (CL, 2 sequences). The gait data is provided as binary silhouette sequences, which are resized to 64×44 pixels.

Legacy protocol retained for baseline-matched comparison. We follow the same closed-set evaluation setup as our previous work [1] to isolate architecture-level improvements. The full 124-subject pool is retained, the gallery uses NM#01-04 from all 11 viewpoints, and the probe set contains NM#05-06, BG#01-02, and CL#01-02. For each probe sample, retrieval is performed over all gallery viewpoints except the same view as the probe. Under this legacy setting, gallery and probe sequences are disjoint at evaluation time, but subject identities are not separated between training and testing. We report Rank-1 per view and cross-view mean Rank-1.

Supplementary subject-disjoint split. To address the generalization concern raised in review, we additionally conduct a subject-disjoint experiment in which subjects 001–074 are used for training and subjects 075–124 are used for testing. The gallery still uses NM#01-04, and the probe still uses NM#05–06, BG#01–02, and CL#01–02, with same-view matches excluded. This split is reported as supplementary evidence and does not replace the protocol-matched comparison above.

Protocol boundary. The legacy setting remains useful for controlled comparison with the previously published Informatica baseline, but it should not by itself be interpreted as evidence of unseen-subject generalization. The additional subject-disjoint split provides a first check of that issue, yet it is still limited to a single split on CASIA-B rather than a full open-set or multi-split evaluation. Accordingly, the manuscript treats the main gains as *within-protocol improvements* and uses the subject-disjoint experiment as supplementary context.

4.2 Implementation details

The model is implemented in PyTorch 2.5.1 and trained on a single NVIDIA GeForce RTX 4090 D GPU. The training configuration is summarized in Table 2.

The PK sampling strategy [18] ensures that each mini-batch contains 8 different identities with 4 sequences each, providing sufficient positive and negative pairs for the batch-hard triplet loss computation. Data augmentation includes random horizontal flipping with 50% probability and random erasing with 30% probability (erasing area ratio 2%–15%), which simulates partial occlusion and enhances robustness. During evaluation, no re-ranking or external post-processing is applied beyond gallery feature averaging and cosine retrieval with L2-normalized embeddings.

Table 3: Training configuration

Parameter	Value
Input preprocessing	Silhouettes resized to 64×44 and normalized to $[0, 1]$
Training frame sampling	30 randomly sampled frames per sequence
Test frame sampling	30 uniformly sampled frames per sequence
Random seed	42
Software environment	Python 3.10.12, PyTorch 2.5.1, CUDA 12.1
Hardware configuration	Single NVIDIA GeForce RTX 4090 D GPU (24 GB VRAM)
Optimizer	Adam
Initial learning rate	1×10^{-4}
Weight decay	5×10^{-4}
LR scheduler	Cosine annealing ($T_{\max} = 500$, $\eta_{\min} = 1 \times 10^{-6}$)
Warmup	Linear warmup for 5 epochs
Batch composition	PK sampler: P=8 identities, K=4 sequences
Effective batch size	32
Frames per sequence	30
Total epochs	500
Dropout rate	0.15
Gradient clipping	max_norm = 5.0
Mixed precision	AMP enabled
Data augmentation	Horizontal flip (50%), Random erasing (30%)
Early stopping	Patience = 100 evaluations (based on NM Rank-1)
Evaluation interval	Every 10 epochs
Best checkpoint rule	Highest NM mean Rank-1 on the evaluation set (Epoch 369 in the current run)
Embedding post-processing	Part-wise feature averaging and L2 normalization; gallery features averaged across NM#01-04 per subject-view
Total training time	$\sim$4.5 hours

4.3 Comparison with the protocol-matched baseline

Table 3 presents the comparison between the proposed model and our previous Informatica baseline [1] under the same legacy protocol.

The proposed model improves all three conditions, with the largest gain on CL (+7.38%). This pattern is consistent with the intended role of channel recalibration and cosine-consistent learning under appearance disturbance.

The CL improvement is important because CL was also the weakest condition in the baseline. In practical terms, this indicates that the proposed modifications improve robustness where silhouette distortion is strongest.

Comparison scope and fairness. In this paper, the primary numerical comparison is made against our previous model under an identical protocol and implementation pipeline. Many public gait baselines report results under different subject splits, preprocessing pipelines, and training recipes. Mixing those numbers in one table as if they were directly comparable would be misleading. We therefore keep the main claim conservative: the proposed changes deliver consistent gains over the protocol-matched baseline, while Table 1 serves only as a broader contextual reference.

Table 4 reports macro-averaged precision, recall, and F1 in addition to Rank-1, showing consistent behavior across classification metrics.

4.4 Per-view rank-1 analysis

Table 5 presents the detailed Rank-1 accuracy for each of the 11 viewpoints under all three conditions. The best model checkpoint (Epoch 369) is used for this evaluation.

Three trends appear in Table 5. First, NM performance is saturated across views (10/11 views at 100%). Second, BG remains stable with small variation across viewpoints. Third, CL shows the largest view sensitivity, with stronger results at side views (90° and 108°) and lower accuracy at 180°. This matches the intuition that clothing alters frontal/rear silhouettes more strongly than lateral leg motion cues.

4.5 Ablation study

To evaluate the contribution of each proposed module, ablation experiments are conducted by removing one module at a time while keeping all other components unchanged. We adopt single-module removal rather than a large combinatorial design so that the marginal effect of each component can be interpreted more directly. Each ablation variant is trained for 300 epochs, and the best checkpoint is selected based on the NM mean Rank-1 accuracy. The results are presented in Table 6.

The ablation results reveal several important findings:

GRU temporal modeling has the largest impact. Removing GRU reduces CL from 98.38% to 72.56% (-25.82),

Table 4: Comparison with the baseline model (Rank-1 accuracy, %)

Model	NM	BG	CL	Params
BFS-CNN-GMP-GRU-MSP [1]	97.00	94.00	91.00	—
Ours	**99.96**	**99.74**	**98.38**	6.83M
Improvement	**+2.96**	**+5.74**	**+7.38**	—

Table 5: Comprehensive evaluation metrics

Condition	Rank-1 (%)	F1-Score	Precision	Recall
NM	99.96	0.9996	0.9996	0.9996
BG	99.74	0.9974	0.9975	0.9974
CL	98.38	0.9833	0.9851	0.9835

indicating that temporal cues remain the dominant signal when spatial appearance is degraded.

GMP is a key bridge to temporal modeling. Removing GMP decreases CL by 20.81 points, suggesting that salient per-frame channel responses are important inputs for GRU.

MSP contributes stable spatial context. Removing MSP causes a 6.15-point drop on CL, showing that multi-scale part-aware features still matter under covariate stress.

BFS and ECA provide complementary gains. Removing BFS and ECA lowers CL by 5.41 and 5.08 points, respectively. ECA offers this gain with only 13 extra parameters.

Module ranking by CL impact: GRU (-25.82) $>$ GMP (-20.81) $>$ MSP (-6.15) $>$ BFS (-5.41) $>$ ECA (-5.08). Under NM, all variants stay above 99%, while under CL the differences become explicit.

4.6 Supplementary subject-disjoint evaluation

To complement the legacy protocol more directly, we additionally run a supplementary subject-disjoint split using subjects 001–074 for training and 075–124 for testing. The gallery/probe composition remains unchanged (gallery: NM#01–04; probe: NM#05–06, BG#01–02, CL#01–02; same-view matches excluded), and the same optimization settings from Table 2 are retained. Because this experiment is intended as a targeted revision check rather than a second full ablation campaign, we report the full model together with two representative variants: w/o ECA and w/o BFS.

The subject-disjoint split is substantially more difficult than the legacy closed-set protocol, especially under CL, but the full model still maintains 97.73%, 87.98%, and 64.00% under NM, BG, and CL, respectively. The resulting pattern is also more nuanced than in Table 6. Removing ECA changes NM and BG only marginally, but reduces CL from 64.00% to 62.18%, suggesting that its main benefit under unseen-subject testing still lies in handling clothing-related appearance disturbance. Removing BFS leads to a different trade-off: BG increases to 90.26%, while NM and CL drop to 97.00% and 63.18%. Taken together, these supplementary results do not suggest that every module improves every condition uniformly; rather, they indicate that the proposed components mainly improve the balance of robustness across covariates, with the clearest effect of ECA appearing under CL.

4.7 Visualization analysis

t-SNE visualization. To reduce cherry-picking, the t-SNE plot is generated from a deterministic subset: the first 15 subject IDs encountered in the evaluation set under the NM condition. All available NM embeddings for these subjects are projected from 256 dimensions into 2D with t-SNE [24]. The resulting clusters are compact and well separated, indicating that the learned representation preserves identity information effectively in the cosine space, although t-SNE remains a qualitative visualization tool rather than a formal statistical test.

Grad-CAM visualization. Gradient-weighted Class Activation Mapping (Grad-CAM) is used to inspect spatial attention during inference. To keep the selection rule deterministic, we visualize the first evaluation sequence and sample 8 uniformly spaced frames from that sequence. High-response regions are concentrated around the lower limbs and torso, which is consistent with gait biomechanics. This provides qualitative support that the model is using motion-relevant body regions rather than background areas.

4.8 Training convergence analysis

The training convergence behavior is summarized from saved checkpoints. Since evaluation is conducted every 10 epochs and model selection is based on NM mean Rank-1, the best-so-far NM trajectory provides a direct view of optimization stability.

As shown in Figure 10, NM performance increases rapidly in the early phase and then enters a stable high-accuracy regime. After epoch 169, the curve remains close to saturation, with the best checkpoint obtained at epoch 369.

Table 6: Per-view Rank-1 accuracy (%)

Condition	0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	Mean
NM	100.0	100.0	100.0	100.0	100.0	100.0	99.60	100.0	100.0	100.0	100.0	99.96
BG	99.59	99.60	99.60	99.59	100.0	100.0	100.0	100.0	99.60	99.18	100.0	99.74
CL	97.56	96.77	97.98	98.79	99.60	100.0	100.0	99.60	98.38	97.14	96.37	98.38

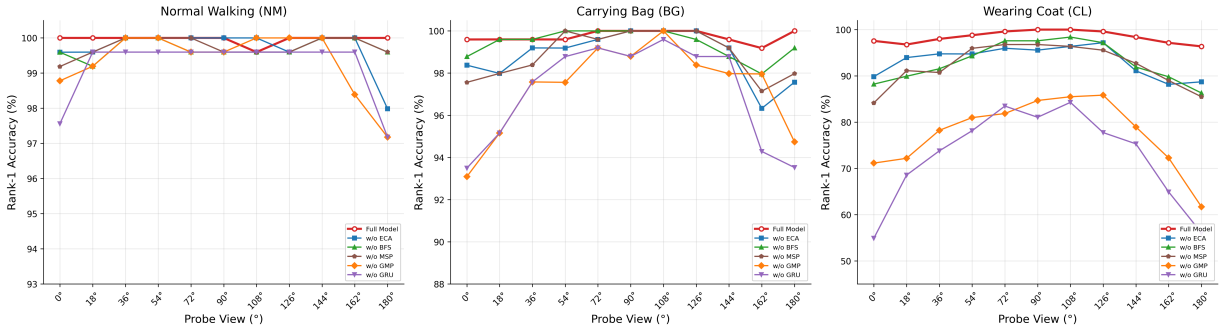

Figure 6: Per-view Rank-1 accuracy (%) across 11 viewpoints under three walking conditions

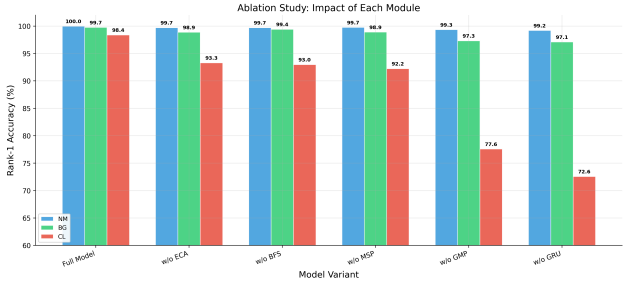

Figure 7: Ablation study results: Rank-1 accuracy (%) of the full model and five ablation variants under NM, BG, and CL conditions.

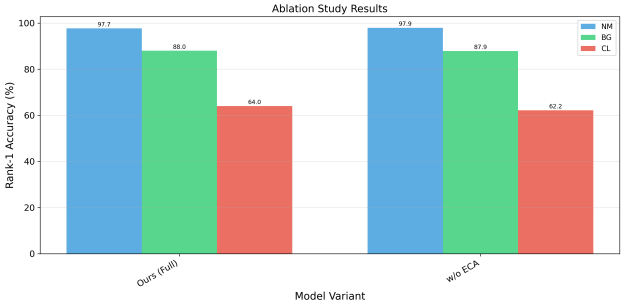

Figure 8: Supplementary subject-disjoint comparison of the full model, w/o ECA, and w/o BFS under NM, BG, and CL conditions.

5 Discussion

Why the gains arise. The empirical pattern is most visible under the CL condition, where the proposed model improves the legacy baseline by 7.38 percentage points. This behavior is consistent with the design motivation of the two new additions. ECA helps suppress less informative or disturbed channels after each backbone stage, which is especially useful when clothing alters the upper-body silhouette. Cosine-consistent learning aligns the training objective and the retrieval metric, reducing the train–test geometry mismatch present in the previous model. The supplementary subject-disjoint split preserves the same qualitative tendency, although the effect becomes more concentrated: ECA mainly benefits CL, while the gains on NM and BG are much smaller.

Relation to representative gait methods. The protocol-aware summary in Table 1 shows that recent gait-recognition models have steadily improved CASIA-B performance through stronger spatial-temporal modeling, larger backbones, or quality-aware processing. Our contribution is different in emphasis: rather than introducing a wholesale backbone replacement, we show that a lightweight attention branch and a cosine-consistent objective can substantially strengthen a protocol-matched baseline with only 13 additional ECA parameters. Because the protocols differ, however, Table 1 should be read as contextual positioning rather than a claim of direct numerical superiority.

Efficiency and deployment perspective. Practical surveillance or intelligent-transportation deployments must balance recognition quality with computational cost. From this perspective, the most attractive property of the proposed enhancement is that the attention branch adds only 13 learnable weights on top of a 6.83M-parameter network, while avoiding expensive transformer-style global attention. Although the present revision does not include a hardware-normalized wall-clock benchmark across external baselines, the negligible parameter increase and the use of a conventional CNN–GRU pipeline preserve a realistic deployment path on single-GPU or edge-server systems.

Protocol boundary and statistical caution. The main

Table 7: Ablation study results (Rank-1 accuracy, %)

Model variant	NM	BG	CL	ΔCL
Full Model	**99.96**	**99.74**	**98.38**	—
w/o ECA	99.71	98.86	93.30	−5.08
w/o BFS	99.71	99.41	92.97	−5.41
w/o MSP	99.74	98.89	92.24	−6.15
w/o GMP	99.34	97.31	77.57	−20.81
w/o GRU	99.19	97.09	72.56	−25.82

Table 8: Supplementary subject-disjoint evaluation on CASIA-B (Rank-1 accuracy, %). Training subjects: 001–074; testing subjects: 075–124.

Model variant	NM	BG	CL	ΔCL
Full Model	97.73	87.98	**64.00**	—
w/o ECA	**97.91**	87.89	62.18	−1.82
w/o BFS	97.00	**90.26**	63.18	−0.82

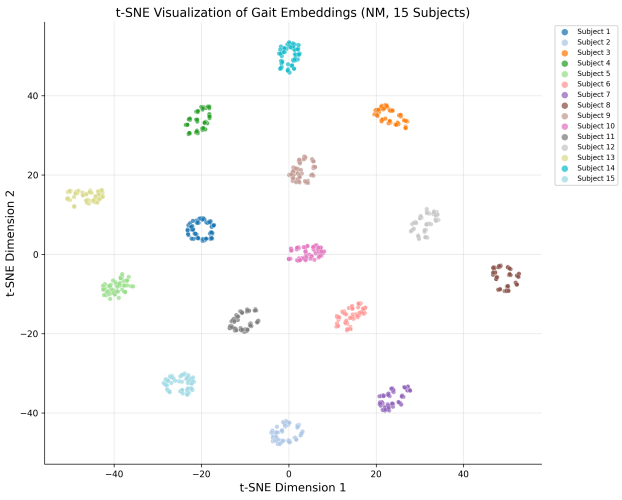

Figure 9: t-SNE visualization of embedding vectors (15 subjects)

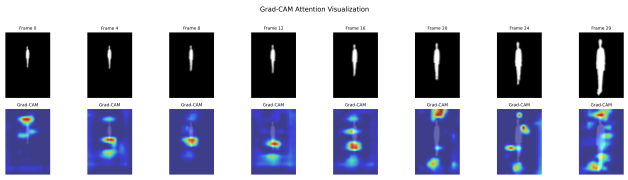

Figure 10: Grad-CAM attention heatmaps on sample frames

Table 9: Training convergence milestones (NM best-so-far Rank-1)

Epoch	NM best-so-far Rank-1 (%)
9	34.87
49	85.02
169	99.19
369	**99.96**

experiments deliberately retain a legacy closed-set protocol for controlled comparison with our previous Informatica paper. This design supports the claim that the proposed modifications improve that protocol-matched baseline. In the revision, we additionally provide a supplementary subject-disjoint split to give a first look at unseen-subject behavior, but that result is still limited to one split on CASIA-B and does not establish open-set generalization. In addition, all gains are reported descriptively from the current training campaign rather than as a multi-seed inferential study, so they should be interpreted as strong empirical margins rather than formal significance claims.

Limitations and ethical considerations. Several limitations remain. First, the evaluation is confined to CASIA-B, an indoor silhouette dataset with controlled viewpoints. Second, silhouette-based recognition may degrade under severe occlusion, poor segmentation, low resolution, or highly unconstrained outdoor settings. Third, even with the added subject-disjoint split, the present evidence is still limited to a single benchmark and a single split, so broader generalization claims remain premature. Beyond technical limitations, gait recognition raises privacy and misuse concerns because it can enable remote identification without active subject cooperation. Any real-world deployment should therefore be constrained by legal compliance, purpose limitation, privacy protection, and human oversight.

6 Conclusion

This work presented an enhanced cross-perspective gait recognition model (BFS-CNN-ECA-GMP-GRU-MSP) built on our earlier BFS-CNN-GMP-GRU-MSP framework. Two changes were introduced: lightweight channel recalibration (ECA) and cosine-consistent metric learning across training and inference.

Under the same legacy CASIA-B protocol retained from

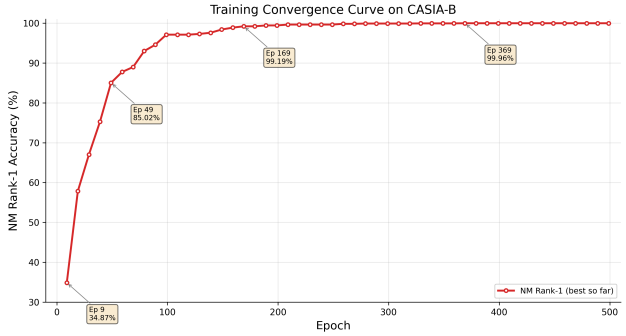

Figure 11: Training convergence curve of the proposed model over 500 epochs

the earlier Informatica study, the model reaches 99.96% (NM), 99.74% (BG), and 98.38% (CL), improving the protocol-matched baseline by +2.96, +5.74, and +7.38 points. A supplementary subject-disjoint split (001–074 for training, 075–124 for testing) further yields 97.73% (NM), 87.98% (BG), and 64.00% (CL). Under this stricter setting, ECA still shows its clearest contribution under the clothing condition, while BFS contributes a different trade-off across NM, BG, and CL. Ablation results under the legacy protocol also show that temporal modeling (GRU/GMP) drives most of the gain, while ECA provides additional robustness at minimal parameter cost.

The main limitation is that, although a supplementary subject-disjoint split is now included, the evidence is still confined to CASIA-B and a single training/testing partition. Future work will therefore prioritize broader subject-disjoint and open-set protocols, followed by validation on larger outdoor benchmarks (e.g., GREW and Gait3D), multi-seed statistical analysis, and deployment-oriented compression or distillation. These next steps are necessary before converting the present within-protocol gain and preliminary unseen-subject evidence into a stronger generalization claim.

Acknowledgment

This work was supported by the Henan Engineering Technology Research Center of Intelligent Transportation Video Image Perception and Recognition, Zhengzhou 451150, China, under Project No. ETRC-250109 (“Key Technologies for Cross-View Robust Gait Recognition in Intelligent Transportation Scenarios”).

Data availability

The CASIA-B dataset is publicly available for academic research from the Institute of Automation, Chinese Academy of Sciences (CASIA). The submission package includes model configuration files, training/evaluation scripts, and figure/table generation scripts used in this study. A reviewer-accessible archive or repository link can be provided in the revision package when permitted by the submission system; otherwise, the source code and checkpoints will be shared upon editorial request and released publicly after the review process.

References

[1] Liu J, Wang W. A Cross-Perspective Gait Recognition Framework Integrating Breadth-First Search and Multi-Scale Feature Map Interaction. *Informatica*, 2022.

[2] Sepas-Moghaddam A, Etemad A. Deep Gait Recognition: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023, 45(1): 264-284.

[3] Fan C, Liang J, Shen C, et al. OpenGait: Revisiting Gait Recognition Towards Better Practicality. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[4] Munusamy V, Senthilkumar S. Emerging Trends in Gait Recognition Based on Deep Learning: A Survey. *PeerJ Computer Science*, 2024, 10: e2158.

[5] Parashar A, Parashar A, Ding W. Deep Learning Pipelines for Recognition of Gait Biometrics with Covariates: A Comprehensive Review. *Artificial Intelligence Review*, 2023, 56(18): 8889-8953.

[6] Khaliluzzaman M, Uddin A, Deb K, et al. Person Recognition Based on Deep Gait: A Survey. *Sensors*, 2023, 23(10): 4875.

[7] Yan S, Hu L, Xueling F. GaitASMS: Gait Recognition by Adaptive Structured Spatial Representation and Multi-Scale Temporal Aggregation. *Neural Computing & Applications*, 2024, 36(13): 7057-7069.

[8] Chao H, He Y, Zhang J, et al. GaitSet: Regarding Gait as a Set for Cross-View Gait Recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, 33(01): 8126-8133.

[9] Fan C, Peng Y, Cao C, et al. GaitPart: Temporal Part-Based Model for Gait Recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020: 14225-14233.

[10] Lin B, Zhang S, Yu X. Gait Recognition via Effective Global-Local Feature Representation and Local Temporal Aggregation. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021: 14648-14656.

[11] Wang M, Guo X, Lin B, et al. DyGait: Exploiting Dynamic Representations for High-performance Gait Recognition. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023: 13424-13433.

[12] Wang Z, Hou S, Zhang M, et al. QAGait: Revisit Gait Recognition from a Quality Perspective. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, 38(6): 5785-5793.

[13] Wang Q, Wu B, Zhu P, et al. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020: 11534-11542.

[14] Hu J, Shen L, Sun G. Squeeze-and-Excitation Networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[15] Yu S, Tan D, Tan T. A Framework for Evaluating the Effect of View Angle, Clothing and Carrying Condition on Gait Recognition. *Proceedings of the 18th International Conference on Pattern Recognition (ICPR)*, 2006: 441-444.

[16] Fan C, Ma J, Jin D, et al. SkeletonGait: Gait Recognition Using Skeleton Maps. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, 38(2): 1662-1669.

[17] Guo M, Xu T, Liu J, et al. Attention Mechanisms in Computer Vision: A Survey. *Computational Visual Media*, 2022, 8(3): 331-368.

[18] Hermans A, Beyer L, Leibe B. In Defense of the Triplet Loss for Person Re-Identification. *arXiv preprint arXiv:1703.07737*, 2017.

[19] Wojke N, Bewley A. Deep Cosine Metric Learning for Person Re-identification. *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018: 748-756.

[20] Luo H, Gu Y, Liao X, et al. Bag of Tricks and a Strong Baseline for Deep Person Re-Identification. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019.

[21] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the Inception Architecture for Computer Vision. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016: 2818-2826.

[22] Yu F, Koltun V. Multi-Scale Context Aggregation by Dilated Convolutions. *International Conference on Learning Representations (ICLR)*, 2016.

[23] Chung J, Gulcehre C, Cho K, et al. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *NIPS Workshop on Deep Learning*, 2014.

[24] Van der Maaten L, Hinton G. Visualizing Data Using t-SNE. *Journal of Machine Learning Research*, 2008, 9(86): 2579-2605.