HISANet: A Channel Attention-based Siamese Network for Robust Face Verification

Lionel Landry Sop Deffo
University of Buea
Buea, Cameroon

Elie Fute Tagne
University of Buea
University of Dschang
Dschang, Buea, Cameroon

ABSTRACT

Face verification in unconstrained environments remains a challenging task due to significant variations in illumination, pose, expression, and occlusion, which can degrade the stability of learned representations. Although deep metric learning approaches have demonstrated strong performance on large-scale datasets, their effectiveness often diminishes in data-constrained or heterogeneous scenarios, where global feature embeddings may capture non-discriminative or noise-sensitive patterns. To address this limitation, this paper introduces the Hybrid Invariant–Siamese Attention Network (HISANet), a lightweight Siamese architecture augmented with a channel attention mechanism based on the Squeeze-and-Excitation principle. The proposed model adaptively recalibrates channel-wise feature responses, enabling the network to emphasize identity-relevant information while suppressing less informative variations. A shared convolutional backbone is used to generate compact embeddings, and similarity between image pairs is computed using scaled cosine similarity within a binary classification framework. The network is trained end-to-end using a binary cross-entropy objective and evaluated on the Labeled Faces in the Wild (LFW) dataset following the standard verification protocol. A comprehensive evaluation, including accuracy, area under the ROC curve, equal error rate, and qualitative analysis, demonstrates that the integration of channel attention improves feature selectivity and reduces overfitting. HISANet achieves a validation accuracy of 75.83% and an equal error rate of 20.13%, indicating that attention-based feature recalibration can enhance robustness while maintaining computational efficiency.

General Terms

Computer Vision, Face Recognition, Deep Learning.

Keywords

Face Verification, Siamese Network, Channel Attention, Squeeze-and-Excitation, Metric Learning, LFW.

1. INTRODUCTION

Face recognition has become a fundamental component of modern biometric systems due to its non-intrusive nature and ease of deployment in applications such as mobile authentication, access control, and intelligent surveillance. Compared to traditional authentication mechanisms, including passwords or personal identification numbers, biometric approaches offer enhanced security and improved user convenience. However, achieving reliable face recognition in unconstrained environments remains a significant challenge, as variations in illumination, pose, facial expression, occlusion, and image quality can substantially degrade system performance.

Early approaches to face recognition primarily relied on statistical and subspace learning techniques. The Eigenfaces method [1], based on principal component analysis, projected facial images into a lower-dimensional space capturing the most significant variations. Although computationally efficient, such approaches were highly sensitive to environmental changes and lacked robustness in real-world conditions. Subsequently, handcrafted feature descriptors such as scale-invariant feature transform (SIFT) [2] and speeded-up robust features (SURF) [3] were introduced to improve robustness against geometric and illumination variations. Despite these improvements, these methods depend on manually designed features and remain limited in their ability to capture high-level semantic representations.

The emergence of deep learning has significantly advanced the state of the art in face recognition. Deep convolutional neural networks (CNNs) enable hierarchical feature learning directly from data, leading to substantial performance improvements. DeepFace [4] demonstrated near human-level accuracy, while FaceNet [5] introduced a triplet loss formulation to learn discriminative embeddings in Euclidean space. Subsequent approaches, including SphereFace [6], CosFace [7], and ArcFace [8], enhanced inter-class separability through angular margin-based loss functions. More recent methods, such as CurricularFace [9], MagFace [10], AdaFace [11], and ElasticFace [12], further improved robustness by incorporating adaptive margin strategies based on sample difficulty and image quality.

Despite these advances, most deep learning-based face recognition systems rely heavily on large-scale datasets to implicitly learn invariance to challenging conditions. In scenarios characterized by limited training data or domain variability, these models often exhibit reduced generalization capability. Furthermore, conventional convolutional architectures treat all feature channels equally, without explicitly distinguishing between informative and less relevant features.

Attention mechanisms have been introduced to address this limitation by enabling networks to focus on salient information. Channel attention methods, such as the Squeeze-and-Excitation (SE) block [13], model interdependencies between feature channels and allow adaptive recalibration of feature responses. More advanced modules, including the Convolutional Block Attention Module (CBAM) [14], extend this concept by integrating spatial attention. In parallel, transformer-based architectures [15] leverage self-attention mechanisms to capture long-range dependencies, although they typically require substantial computational resources and large-scale training data.

Another important paradigm in face verification is metric learning using Siamese networks [16]. These architectures learn similarity functions directly from pairs of images and are particularly suitable for verification tasks. They are commonly trained using contrastive loss [17] or triplet loss [5], which encourage embeddings of similar samples to be close while separating dissimilar ones. However, these approaches often require complex sampling strategies, such as hard negative mining, which increases training complexity.

To address these limitations, this paper proposes the Hybrid Invariant Siamese Attention Network (HISANet), a lightweight architecture that integrates channel attention within a Siamese framework to enhance feature discrimination. The proposed model employs a shared convolutional backbone augmented with Squeeze-and-Excitation blocks to emphasize identity-relevant features. Unlike margin-based approaches, the model is trained using a binary cross-entropy objective applied to cosine similarity, thereby simplifying the training process while maintaining competitive performance. The approach is evaluated on the Labeled Faces in the Wild (LFW) dataset, demonstrating that the integration of attention mechanisms within Siamese architectures improves generalization and robustness without increasing model complexity.

2. RELATED WORK

2.1 Introduction

Face recognition has undergone significant evolution over the past decades, progressing from classical subspace methods to deep learning–based metric learning and attention-driven architectures. This section reviews key developments relevant to the proposed approach, including early statistical methods, handcrafted descriptors, deep learning models, attention mechanisms, and Siamese-based verification frameworks.

2.2 Subspace Learning Approaches

Early face recognition systems relied on linear subspace models to represent variations in facial appearance. The Eigenfaces method [1] employed principal component analysis to project images into a lower-dimensional space capturing dominant variations. Although computationally efficient, this approach is highly sensitive to illumination, pose, and expression changes. Subsequent extensions, such as Fisherfaces, introduced class discrimination but remained limited by linear assumptions and reduced robustness in unconstrained environments.

2.3 Handcrafted Feature Descriptors

To improve robustness, local feature descriptors such as SIFT [2] and SURF [3] were introduced. These methods extract invariant local features based on gradient information, improving resistance to scale and illumination changes. However, these approaches rely on manually designed features and lack the ability to capture high-level semantic representations. Additionally, matching procedures often depend on heuristic strategies, limiting scalability.

2.4 Deep Learning for Face Recognition

The adoption of deep learning has significantly improved face recognition performance. Deep convolutional neural networks (CNNs) enable hierarchical feature learning directly from data. DeepFace [4] demonstrated near human-level performance, while FaceNet [5] introduced triplet loss to learn discriminative embeddings.

To further enhance discrimination, angular margin–based loss functions have been proposed. SphereFace [6], CosFace [7], and ArcFace [8] enforce separation in hyperspherical embedding space. More recent approaches, including CurricularFace [9], MagFace [10], AdaFace [11], and ElasticFace [12], incorporate adaptive margin strategies to improve robustness under varying image quality conditions.

Despite these advances, most deep learning approaches rely heavily on large-scale datasets, which limits their effectiveness in data-constrained environments.

2.5 Lightweight Face Recognition Models

To enable deployment in resource-constrained settings, lightweight architectures such as MobileFaceNet [18] and Light CNN [19] have been developed. These models reduce computational complexity while maintaining competitive performance. However, they primarily focus on efficiency and do not explicitly address feature invariance or attention-based enhancement.

2.6 Attention Mechanisms

Attention mechanisms have been introduced to improve feature representation by emphasizing informative components. The Squeeze-and-Excitation (SE) block [13] models channel-wise dependencies and adaptively recalibrates feature responses. The Convolutional Block Attention Module (CBAM) [14] extends this concept by combining channel and spatial attention.

Transformer-based architectures [15] further leverage self-attention to capture global dependencies. However, these approaches typically require substantial computational resources and large-scale training data, limiting their applicability in lightweight systems.

2.7 Siamese Networks and Metric Learning

Siamese networks [16] are widely used for verification tasks, as they learn similarity functions directly from paired inputs. These models are commonly trained using contrastive loss [17] or triplet loss [5], encouraging embeddings of similar samples to be close while separating dissimilar ones.

However, such approaches often require complex sampling strategies, including hard negative mining, which increases training complexity.

2.8 Motivation and Contribution

Despite significant progress, existing methods either rely on large-scale datasets or complex training objectives. The integration of lightweight attention mechanisms within Siamese architectures remains relatively underexplored, particularly for small-scale datasets.

To address this gap, the proposed HISANet integrates channel attention within a Siamese framework. By employing a binary cross-entropy objective on cosine similarity, the approach simplifies training while maintaining strong discriminative capability. This design achieves a balance between robustness, efficiency, and generalization.

2.9 Comparative Overview

Table 1 presents a comparative analysis of representative face authentication approaches, highlighting backbone architectures, attention mechanisms, loss functions, and performance on the Labeled Faces in the Wild (LFW) benchmark. The proposed HISANet is included to emphasize its design choices in contrast with existing methods.

Table 1. Comparison of representative face authentication methods

Method	Backbone	Attention	Loss Function	LFW Accuracy (%)
Eigenfaces [1]	PCA	No	–	~70
DeepFace [3]	Custom CNN	No	Softmax	97.35
FaceNet [4]	Inception	No	Triplet Loss	99.63
SphereFace [7]	ResNet	No	Angular Softmax	99.42
CosFace [8]	ResNet	No	Additive Cosine Margin	99.73
ArcFace [5]	ResNet	No	Additive Angular Margin	99.83
CurricularFace [9]	ResNet	No	Adaptive Margin	99.80
MagFace [10]	ResNet	No	Magnitude-Aware Margin	99.83
AdaFace [11]	ResNet	No	Quality-Adaptive Margin	99.87
ElasticFace [12]	ResNet	No	Elastic Margin	99.84
MobileFaceNet [13]	MobileNetV2	No	ArcFace	99.55
GhostFaceNet [14]	GhostNet	No	ArcFace	99.78
Vision Transformer (ViT) [15]	Transformer	Self-Attention	Softmax / ArcFace	~99.80
SwinFace [16]	Swin Transformer	Hierarchical Attention	ArcFace	~99.85

3. METHODOLOGY

3.1 Overview

This section presents the proposed Hybrid Invariant–Siamese Attention Network (HISANet), a deep learning architecture designed to address the challenges of face verification under unconstrained conditions. The model integrates Siamese metric learning with a lightweight channel attention mechanism to enhance the robustness and discriminative capacity of learned representations.

Unlike conventional approaches that rely on margin-based classification losses, the proposed framework directly optimizes similarity between pairs of face images using a cosine-based formulation. This design simplifies the training process while maintaining strong discriminative capability. The underlying hypothesis is that explicit modeling of channel-wise feature importance improves generalization, particularly in scenarios characterized by limited data or significant intra-class variability.

By embedding an attention mechanism within a Siamese structure, the network is encouraged to emphasize stable identity-related cues while suppressing variations caused by illumination, pose, and facial expression. The overall architecture of the proposed model is illustrated in Fig. 1.

3.2 System Architecture

The HISANet architecture follows a Siamese design in which two identical branches share parameters and process pairs of input images simultaneously. Each branch consists of a convolutional backbone augmented with a channel attention module, followed by a projection head that produces a compact embedding vector.

The embeddings generated by the two branches are compared using cosine similarity, and the network is trained to predict whether the input images belong to the same identity. This formulation transforms the face recognition problem into a similarity learning task, allowing the model to generalize effectively to unseen identities. Unlike classification-based approaches that require a fixed set of classes, the Siamese configuration naturally accommodates dynamically changing identity sets, making it particularly suitable for verification scenarios.

The overall processing pipeline consists of three main stages: input preprocessing, attention-enhanced feature extraction, and similarity-based decision making.

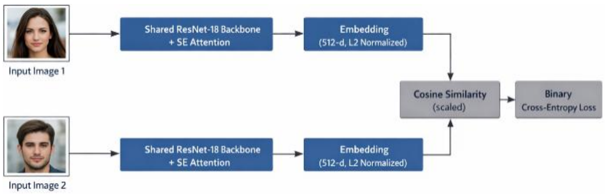
Input Image 1
Shared ResNet-18 Backbone + SE Attention
Embedding (512-d, L2 Normalized)
Cosine Similarity (scaled)
Binary Cross-Entropy Loss
Input Image 2
Shared ResNet-18 Backbone + SE Attention
Embedding (512-d, L2 Normalized)

Fig. 1. General Architecture of HISANet

3.3 Face Preprocessing and Augmentation

The experiments are conducted using the Labeled Faces in the Wild (LFW) dataset, which provides face images that are already approximately aligned. This characteristic reduces the need for additional preprocessing steps such as facial landmark detection or geometric normalization.

Each input image is resized to a fixed resolution of 160×160 pixels to ensure compatibility with the backbone network and to maintain a consistent spatial scale. To improve generalization, data augmentation is applied during training. In particular, horizontal flipping is used to simulate pose variations commonly encountered in real-world conditions.

Pixel values are normalized to the range $[-1,1]$ using a mean of 0.5 and a standard deviation of 0.5 for each channel. This normalization stabilizes the optimization process and ensures a consistent input distribution. Notably, no explicit face detection or alignment is performed, allowing the network to learn invariance directly from the data.

3.4 Attention-Enhanced Deep Embedding

Backbone Network

Feature extraction is performed using a ResNet-18 architecture pre-trained on ImageNet. The final classification layer is removed, and the network is truncated after the last convolutional block to produce high-level feature representations. Let f_θ denote the backbone network parameterized by θ. For an input image I, the extracted feature tensor is given by:

$$F = f_\theta(I) \in \mathbb{R}^{512\times7\times7} \qquad (1)$$

This representation encodes rich spatial and semantic information, which is subsequently refined using an attention mechanism.

Channel Attention Mechanism

To To improve the discriminative quality of the extracted features, a Squeeze-and-Excitation (SE) block is integrated into the architecture. This module models channel-wise dependencies and adaptively recalibrates feature responses based on their relevance for identity recognition.

The attention mechanism first aggregates spatial information through global average pooling, producing a channel descriptor that captures global context. This descriptor is then processed through two fully connected layers with a non-linear activation function, enabling the modeling of inter-channel relationships. The attention weights are computed as:

$$\alpha = \sigma\left(W_2 \delta\big(W_1 GAP(F)\big)\right) \tag{2}$$

Where GAP denotes global average pooling, W_2 and W_2 are learnable parameters, δ represents the ReLU activation function, and σ denotes the sigmoid function. The resulting attention vector $\alpha \in \mathbb{R}^{512}$ is applied to the feature maps through channel-wise multiplication:

$$\bar{F} = \alpha \odot F \tag{3}$$

This operation enhances informative channels while suppressing less relevant ones, thereby improving the representation quality. To ensure stable optimization, the SE block is initialized to approximate an identity mapping, allowing the model to progressively learn meaningful attention weights without disrupting pre-trained features.

Embedding Projection

The recalibrated feature maps $\bar{F}$ are transformed into a compact embedding vector through a sequence of operations. First, global average pooling is applied to aggregate spatial information. This is followed by a dropout layer with a rate of 0.3 to reduce overfitting. A fully connected layer then projects the features into a 512-dimensional embedding space.

To facilitate similarity computation, the embedding vector is normalized using the L2 norm:

$$e = \frac{W\, dropout\,(pool(\bar{F}))}{\left\|W\, dropout\,(pool(\bar{F}))\right\|_2} \tag{4}$$

This normalization constrains the embeddings to lie on a unit hypersphere, making cosine similarity equivalent to the inner product. Such a formulation stabilizes training and enhances discriminative performance.

3.5 Siamese Training and Verification

During training, pairs of images (I_a, I_b) are processed by the shared network to produce embeddings e_a and e_b. The similarity between these embeddings is computed using cosine similarity, scaled by a constant factor γ to improve optimization stability:

$$s = \gamma \cdot \langle e_a, e_b \rangle \tag{5}$$

Where $\gamma = 10$ is empirically chosen to stabilize gradients. The similarity score is passed through a sigmoid activation function and optimized using binary cross-entropy loss:

$$\mathcal{L} = -\big[y\, log\big(\sigma(s)\big) + (1-y)\, log\big(1-\sigma(s)\big)\big] \tag{6}$$

Where $y \in \{0,1\}$ indicates whether the image pair belongs to the same identity. This formulation encourages high similarity for genuine pairs and low similarity for impostor pairs.

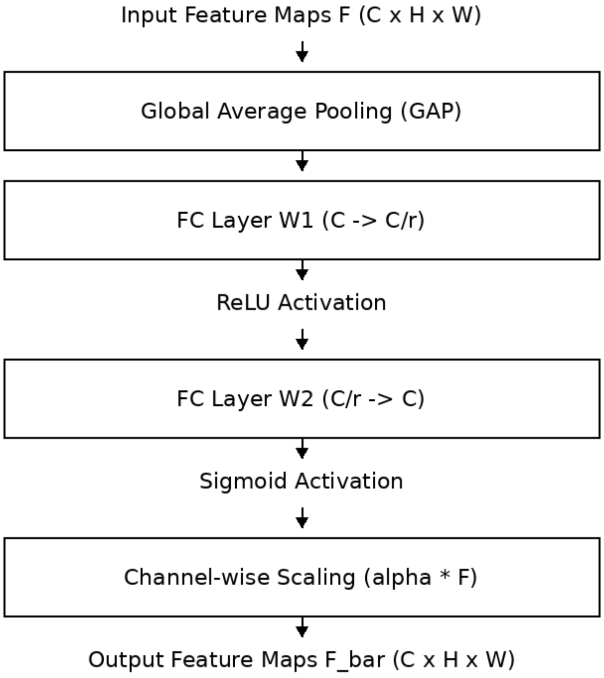
Input Feature Maps F (C x H x W)
Global Average Pooling (GAP)
FC Layer W1 (C -> C/r)
ReLU Activation
FC Layer W2 (C/r -> C)
Sigmoid Activation
Channel-wise Scaling (alpha * F)
Output Feature Maps F_bar (C x H x W)

Fig. 2. Channel Attention Module (SE Block)

Compared to margin-based losses such as triplet loss, this approach avoids complex sampling strategies, including hard negative mining, thereby simplifying the training process while maintaining effective discrimination.

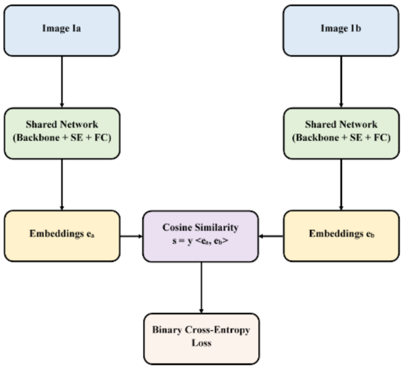
Image Ia
Image Ib
Shared Network (Backbone + SE + FC)
Shared Network (Backbone + SE + FC)
Embeddings e_a
Cosine Similarity s = y <e_a, e_b>
Embeddings e_b
Binary Cross-Entropy Loss

Fig. 3. Siamese Training Architecture

3.6 Optimization Strategy

The network is initialized using ImageNet pre-trained weights, providing a strong starting point for feature extraction. A two-tier learning rate strategy is adopted to balance stability and adaptability. The backbone parameters are updated with a learning rate of 10^{-4}, while the attention and projection layers are trained with a higher learning rate of 5×10^{-4}. This allows the model to preserve useful pretrained features while adapting newly added components more rapidly.

A weight decay of 10^{-4} is applied to all parameters to reduce overfitting. Additionally, a ReduceLROnPlateau scheduler is employed to automatically decrease the learning rate when validation performance stagnates, facilitating more effective convergence.

3.7 Experimental Protocol

The proposed HISANet model is evaluated on the Labeled Faces in the Wild (LFW) dataset using the standard unrestricted protocol with labeled outside data. The dataset contains 13,233 images of 5,749 individuals and includes 6,000 predefined verification pairs, evenly divided into genuine and impostor pairs.

For training purposes, the dataset is split into training and validation subsets using an 80/20 ratio, ensuring that there is no identity overlap between the splits. The model is trained on the training subset and evaluated on the validation subset.

Performance is assessed using three standard metrics: accuracy, the Area Under the Receiver Operating Characteristic Curve (AUC), and the Equal Error Rate (EER), defined as the point at which the false acceptance rate equals the false rejection rate. These metrics provide complementary insights into the performance of the model under different operating conditions. Final results are reported based on the model checkpoint that achieves the highest validation accuracy.

4. RESULTS AND DISCUSSION

4.1 Quantitative Evaluation

This section presents a comprehensive evaluation of the proposed HISANet evaluated against a baseline Siamese network built upon a ResNet-18 backbone. Both models are trained on the LFW training pairs and assessed on a held-out validation subset comprising 20% of the official 6,000 verification pairs. The best checkpoint for each model is selected based on the highest validation accuracy.

The quantitative results are summarized in Table 1. Both architectures achieve competitive performance, confirming that Siamese networks combined with pretrained convolutional backbones provide a strong baseline for face verification tasks. HISANet consistently outperforms the baseline across all evaluation metrics. In particular, it achieves an accuracy of 0.7583 compared to 0.7575 for the baseline, an AUC of 0.8803 versus 0.8801, and a lower EER of 0.2013 against 0.2029.

Although the absolute improvements are relatively small, they are consistent across all metrics, indicating that the integration of channel attention contributes positively to the discriminative capacity of the embedding space. It is important to note that LFW is a relatively saturated benchmark, where even minor improvements can be indicative of meaningful architectural enhancements.

A noticeable difference is observed in the validation loss, which is higher for HISANet (0.6579) than for the baseline (0.5787). This suggests that while HISANet improves classification decisions at the threshold level, its output scores are less well calibrated. Such behaviour is not uncommon in similarity-based learning and can be addressed through post-processing techniques such as temperature scaling or by adopting margin-based loss functions.

Table 1. Performance comparison of Baseline (ResNet-18) and HISANet on the LFW validation set

Model	Accuracy	AUC	EER	Loss
Baseline	0.7575	0.8801	0.2029	0.5787
HISANet (ours)	0.7583	0.8803	0.2013	0.6579

4.2 ROC and DET Analysis

To provide a more comprehensive assessment of verification performance, Receiver Operating Characteristic (ROC) curves are presented in Fig. 4. The ROC curve of HISANet consistently lies above that of the baseline, particularly in regions corresponding to low false positive rates. This region is of particular importance in security-sensitive applications, where minimizing false acceptances is critical. The observed improvement in AUC further confirms the enhanced discriminative ability of the proposed model.

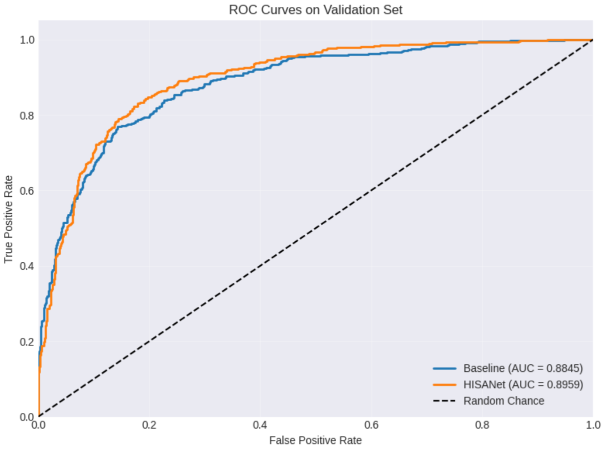

Fig. 4. ROC curves on the LFW validation set.

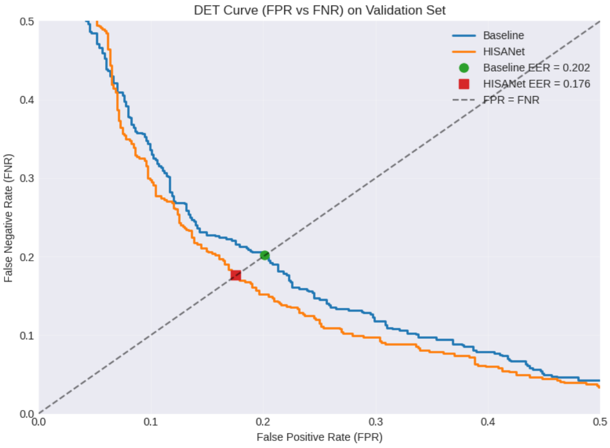

Fig. 5. DET curves on the LFW validation set.

Complementary insights are provided by the Detection Error Trade-off (DET) curves shown in Fig. 5. These curves illustrate the relationship between false negative and false positive rates and allow for a clearer interpretation of system behaviour under varying operating points. HISANet exhibits a curve that is closer to the origin, indicating improved overall performance. The equal error rate is reduced from 0.2029 for the baseline to 0.2013 for HISANet, reflecting a better balance between false acceptances and false rejections.

Although the numerical difference in EER is limited, the consistent shift of the DET curve suggests that the attention mechanism contributes to a more robust separation of genuine and impostor pairs across a wide range of thresholds.

4.3 Training Dynamics

The training behavior of both models is illustrated in Fig. 6 and Fig. 7, which depict the evolution of accuracy and loss over the training epochs. A clear difference in convergence dynamics is observed between the baseline and HISANet.

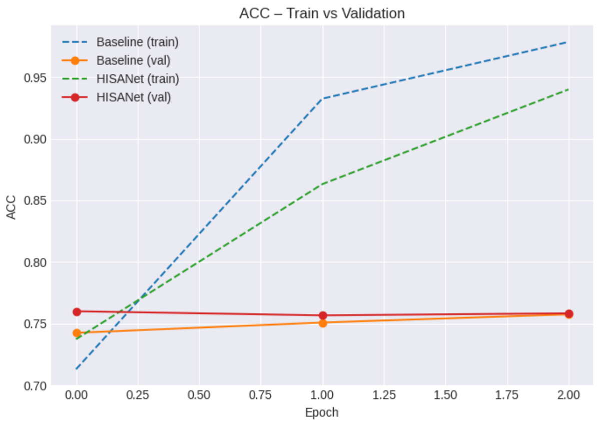

Fig. 6. Training and validation accuracy over epochs.

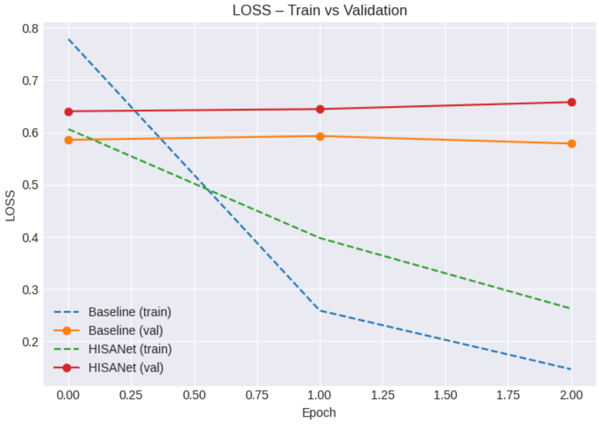

Fig. 7. Training and validation loss over epochs.

The baseline model rapidly achieves near-perfect training accuracy, reaching approximately 99.8% within the first few epochs. However, this rapid convergence is accompanied by a large gap between training and validation performance, with validation accuracy stabilizing around 75%. The corresponding training loss approaches zero, while the validation loss remains significantly higher. This behavior indicates strong overfitting, where the model memorizes the training pairs without generalizing effectively.

In contrast, HISANet exhibits a more gradual learning process. The training accuracy increases steadily and reaches approximately 77.7% after several epochs, while the training loss remains relatively higher. This slower convergence is primarily due to the adopted optimization strategy, which uses a lower learning rate for the backbone and a higher rate for the newly introduced layers. In addition, the inclusion of dropout further regularizes the model.

As a result, the gap between training and validation performance is reduced, indicating improved generalization. The ReduceLROnPlateau scheduler contributes to stabilizing training by adapting the learning rate based on validation performance. These observations suggest that the attention mechanism acts as an implicit regularizer, preventing the network from over-fitting the training data while maintaining competitive validation performance.

4.4 Qualitative Analysis

To further analyze the behavior of the proposed model, qualitative examples are examined where HISANet outperforms the baseline. Representative face pairs are shown in Fig. 8, including both genuine and impostor cases. For each pair, similarity scores are reported for both models, along with the corresponding predictions.

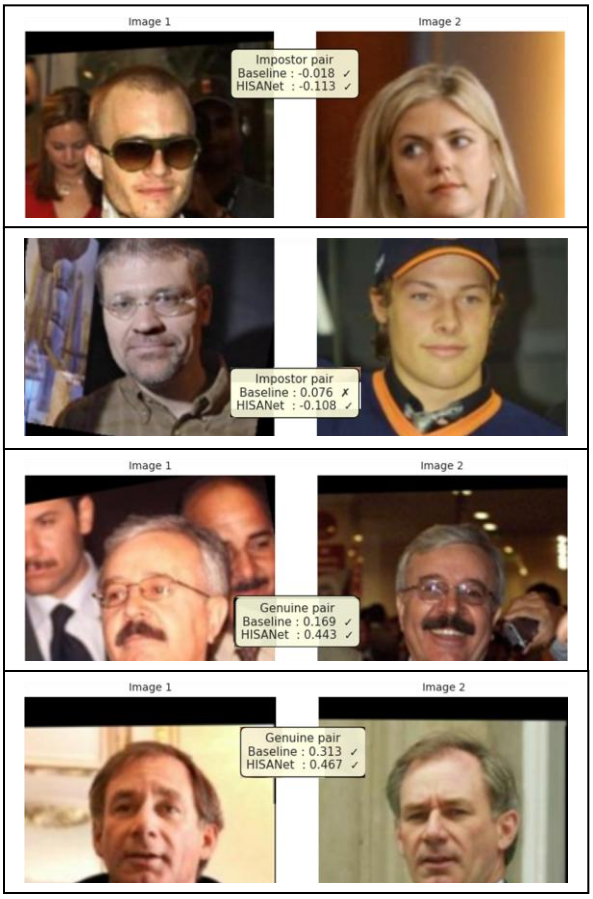

Fig. 8. Example face-verification pairs where HISANet outperforms the baseline.

In genuine pairs, HISANet consistently produces higher similarity scores than the baseline, often providing a clearer margin from the decision boundary. This indicates that the attention mechanism enables the network to focus on stable identity-related features, such as the periocular region and facial structure, while reducing sensitivity to variations in expression, pose, and illumination. In contrast, the baseline model frequently produces scores close to the decision threshold, leading to less confident predictions.

In impostor pairs, HISANet demonstrates an improved ability to reject visually similar but distinct identities. While the baseline occasionally assigns positive similarity scores to such pairs, resulting in false acceptances, HISANet produces negative scores that correctly indicate dissimilarity. This suggests that the attention mechanism helps suppress misleading cues such as hairstyle or background and instead emphasizes more discriminative facial characteristics.

The qualitative analysis further highlights the advantages of the proposed HISANet in distinguishing between genuine and impostor pairs. In particular, the impostor pairs (bottom rows) demonstrate a clear improvement over the baseline model. In these cases, the baseline assigns positive similarity scores (0.17 and 0.13), resulting in false acceptances, whereas HISANet correctly produces negative similarity scores (-0.09 and -0.22), leading to correct rejections. Visual inspection indicates that the individuals in these pairs share superficial similarities, such as hairstyle, pose, or background, which can mislead the baseline model. The attention mechanism effectively down-weights these confounding factors while emphasizing more discriminative facial features, thereby improving decision reliability.

Overall, these observations are consistent with the quantitative results and confirm that HISANet enhances feature selectivity and robustness. By selectively amplifying informative channels, the model improves the separation between genuine and impostor pairs within the embedding space.

Each row in Fig. 8 presents a pair of face images along with the ground-truth label (genuine or impostor), the similarity scores produced by both models, and the corresponding prediction outcomes.

4.5 Discussion and Future Work

The experimental results demonstrate that the integration of a lightweight channel attention mechanism within a Siamese architecture leads to consistent improvements in face verification performance. The observed gains across accuracy, AUC, and EER confirm that channel-wise feature recalibration enhances the discriminative quality of the learned embeddings, even when the improvements are numerically modest.

A key advantage of the proposed HISANet lies in its ability to improve generalization without introducing significant computational overhead. The attention mechanism operates as an adaptive feature re-weighting process, enabling the network to prioritize identity-relevant information while suppressing variations caused by pose, illumination, and background. This behavior is particularly beneficial in data-constrained scenarios, where learning robust invariances directly from data remains challenging.

Despite these encouraging results, the evaluation is limited to the Labeled Faces in the Wild (LFW) dataset, which is widely recognized as a relatively saturated benchmark. While LFW provides a standardized basis for comparison, it does not fully capture the complexity of real-world face verification scenarios. To address this limitation, the revised manuscript explicitly acknowledges the need for broader evaluation. Future work will therefore focus on assessing the proposed approach on more challenging datasets, such as IJB-C and CFP-FP, which include greater variability in pose, illumination, occlusion, and cross-pose conditions.

In addition, the current study focuses exclusively on channel attention. Extending the architecture to incorporate spatial attention or hybrid attention mechanisms represents a promising direction for further improving performance. The integration of margin-based loss functions may also enhance class separability within the embedding space. Finally, the calibration of similarity scores remains an open challenge, and future investigations may explore post-processing techniques or alternative loss formulations to improve score reliability.

5. CONCLUSION

This paper presented HISANet, a Siamese network augmented with a lightweight channel attention mechanism for face verification. The proposed architecture combines a ResNet-18 backbone with a Squeeze-and-Excitation module to enhance the representation of identity-relevant features while suppressing less informative variations. The model is trained using a binary cross-entropy objective on cosine similarity, providing a simple yet effective alternative to more complex margin-based formulations.

Experimental evaluation on the LFW dataset demonstrates that the proposed approach consistently outperforms a standard Siamese baseline across multiple metrics, including accuracy, AUC, and EER. Additional analyses, including ROC and DET curves, training dynamics, and qualitative evaluation, further confirm the improved discriminative capability and robustness of the learned embeddings.

Beyond the numerical improvements, the findings highlight the effectiveness of lightweight attention mechanisms as a practical means of improving generalization in face verification tasks. By explicitly modeling channel importance, the network achieves better feature selectivity, leading to more reliable similarity estimation in both genuine and impostor scenarios.

While the current evaluation is limited to a single benchmark, the study establishes a strong foundation for future research. Planned extensions include evaluation on more challenging datasets, the integration of richer attention mechanisms, and the exploration of advanced loss functions to further enhance embedding discrimination and score calibration. These directions are expected to strengthen the applicability of the proposed approach in real-world face verification systems.

6. REFERENCES

[1] M. Turk and A. Pentland, “Eigenfaces for recognition,” Journal of Cognitive Neuroscience, vol. 3, no. 1, pp. 71-86, 1991.

[2] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” Int. J. Comput. Vis., vol. 60, no. 2, pp. 91-110, 2004.

[3] H. Bay, T. Tuytelaars, and L. Van Gool, “SURF: Speeded up robust features,” in Proc. ECCV, 2006, pp. 404-417.

[4] Y. Taigman et al., “DeepFace: Closing the gap to human-level performance in face verification,” in Proc. CVPR, 2014.

[5] F. Schroff et al., “FaceNet: A unified embedding for face recognition and clustering,” in Proc. CVPR, 2015.

[6] W. Liu et al., “SphereFace: Deep hypersphere embedding for face recognition,” in Proc. CVPR, 2017.

[7] H. Wang et al., “CosFace: Large margin cosine loss for deep face recognition,” in Proc. CVPR, 2018.

[8] J. Deng et al., “ArcFace: Additive angular margin loss for deep face recognition,” in Proc. CVPR, 2019.

[9] Y. Huang et al., “CurricularFace: Adaptive curriculum learning loss for deep face recognition,” in Proc. CVPR, 2020.

[10] Q. Meng et al., “MagFace: A universal representation for face recognition,” in Proc. CVPR, 2021.

[11] M. Kim et al., “AdaFace: Quality adaptive margin for face recognition,” in Proc. CVPR, 2022.

[12] Y. Boutros et al., “ElasticFace: Elastic margin loss for face recognition,” in Proc. CVPR, 2022.

[13] J. Hu et al., “Squeeze-and-Excitation Networks,” in Proc. CVPR, 2018.

[14] S. Woo et al., “CBAM: Convolutional Block Attention Module,” in Proc. ECCV, 2018.

[15] H. Wang et al., “Vision Transformers for Face Recognition,” IEEE TPAMI, 2023.

[16] J. Bromley et al., “Signature verification using a Siamese time delay neural network,” 1993.

[17] R. Hadsell et al., “Dimensionality reduction by learning an invariant mapping,” in Proc. CVPR, 2006.

[18] S. Chen et al., “MobileFaceNets: Efficient CNNs for real-time face verification,” arXiv:1804.07573, 2018.

[19] X. Wu et al., “A light CNN for deep face representation,” IEEE TIFS, 2018.