Jurnal Ilmu Komputer dan Informasi (Journal of Computer Science and Information)
19/1 (2026), 89-97. DOI: http://dx.doi.org/10.21609/jiki.v19i1.1627

Face Gender Recognition Optimization Using VGG-16 With Integration of Spatial Attention Block and Channel Attention Block

Tio Dharmawan, Leonardus Virmus Danar Kusuma Putra, Muhamad Arief Hidayat, Stanislaus Jiwandana Pinasthika

Faculty of Computer Science, Universitas Jember, Jember, Indonesia

E-mail: tio.pssi@unej.ac.id, leonvirmus17@gmail.com, arief.hidayat@unej.ac.id, stanislausjp@unej.ac.id

Abstract

Face gender recognition plays a critical role in applications such as security systems, personalized services, and human-computer interaction. Although VGG-16 is commonly used in this domain, it struggles to retain important spatial information under varying lighting conditions, facial expressions, and viewing angles. This study enhances the VGG-16 model by integrating the Convolutional Block Attention Module (CBAM), which consists of spatial and channel attention mechanisms. Several training scenarios were explored, including applying CBAM to all convolutional blocks and fine-tuning blocks 2 to 5. Experiments conducted on the Labeled Faces in the Wild (LFW) Gender dataset showed a notable improvement in performance. The best configuration achieved an accuracy of 91.78%, outperforming the baseline model (82.13%–88.72%). Other evaluation metrics such as Precision, Recall, and F1-Score also improved, confirming the effectiveness of attention mechanisms in enhancing feature extraction and classification accuracy in face gender recognition tasks.

Keywords: *Face Gender Recognition, VGG-16, Spatial Attention Block, Channel Attention Block, Fine-tuning*

1. Introduction

Face gender recognition refers to the process of automatically identifying an individual's gender based on their facial image [1], [2], [3]. This task has garnered significant attention due to its wide range of applications in various domains, including intelligent surveillance systems, targeted advertising, access control, demographic data collection, and human-computer interaction [4]. Accurate gender recognition can enhance the adaptability of AI systems, enabling more personalized and context-aware responses [5]. However, despite its potential, gender classification based on facial imagery remains a challenging problem due to numerous variations in real-world scenarios.

Several intrinsic and extrinsic factors affect the performance of facial gender recognition systems [6]. These include variations in lighting conditions, facial expressions, head pose, aging, and occlusion (e.g., glasses, hair, masks) [7]. Such variability can obscure gender-specific features, making it difficult for recognition models to generalize effectively across different datasets or environments. The need for robust models that can handle such inconsistencies is therefore critical [8].

Historically, traditional machine learning approaches were employed for this task. Techniques such as Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), and Support Vector Machines (SVM) were commonly used to extract and classify handcrafted features like texture, shape, and color histograms [9], [10], [11]. While these methods performed reasonably well under controlled conditions, they often failed to generalize in unconstrained environments due to their limited capacity to model high-level abstractions and complex visual patterns [12].

With the advent of deep learning, especially Convolutional Neural Networks (CNNs), face gender recognition has seen significant progress [13]. CNNs offer the ability to automatically learn hierarchical and discriminative features directly from raw image pixels without relying on manual feature engineering [14]. This has allowed them to outperform classical approaches, particularly when trained on large and diverse datasets [15]. In addition, the use of transfer learning with pre-trained CNN architectures such as VGGNet, ResNet, GoogLeNet, and EfficientNet has enabled researchers to achieve high accuracy even with limited training data [16].

Among these, the VGG-16 model remains a widely adopted architecture due to its simple yet effective design for visual recognition tasks. VGG-

16 comprises 13 convolutional layers and 3 fully connected layers, forming a deep network capable of capturing rich visual patterns. However, VGG-16 also has limitations. Its architecture lacks an inherent mechanism for focusing on salient regions in an image, which can lead to suboptimal learning in complex scenarios. Furthermore, its high parameter count and computational cost make it less efficient for deployment in real-time or resource-constrained environments [17].

To address these challenges, this study proposes a modified version of the VGG-16 architecture by incorporating attention mechanisms, namely the Spatial Attention Block (SAB) and the Channel Attention Block (CAB) [18]. The SAB enables the network to focus on spatially important regions of the face (e.g., eyes, nose, and mouth), while the CAB allows it to selectively emphasize feature maps that carry gender-discriminative information [19]. These modules are collectively referred to as the Convolutional Block Attention Module (CBAM). Despite introducing additional parameters, CBAM substantially improve spatial anda channel-wise feature refinement, and the computational overhead remains minimal.This make CBAM an efficient enhancement for strengthening representation quality [20].

Previous research has shown the effectiveness of combining attention mechanisms with deep CNNs [21]. For instance, studies have indicated that two-stage training strategies where earlier layers are initially frozen and deeper layers are later fine-tuned can improve performance on gender classification tasks. One such study reported an accuracy of 91.44% on the LFW dataset using a modified VGG-16 with attention modules. However, other findings have also pointed out potential drawbacks of attention-based models, including parameter redundancy, increased training time, and susceptibility to overfitting, particularly when applied indiscriminately.

Building upon these insights, this research enhances the VGG-16 model by integrating CBAM components in a systematic and controlled manner [22]. The training process is divided into two key phases: (1) training the dense layers while keeping the convolutional base frozen, and (2) fine-tuning the last few convolutional blocks embedded with SAB and CAB. This approach enables the model to retain general visual features while learning task-specific discriminative features for gender classification [23]. The effectiveness of the proposed method is evaluated using several performance metrics, including accuracy, precision, recall, and F1-score [24].

This study aims to contribute to the ongoing development of robust gender recognition systems by offering a lightweight yet accurate architecture that can adapt to real-world conditions. By leveraging spatial and channel attention mechanisms in conjunction with transfer learning, the model aspires to improve recognition performance while minimizing overfitting. The proposed method holds promise not only for gender classification but also for broader facial attribute analysis tasks in unconstrained environments.

2. Related Work

Face gender recognition is a research area that aims to determine an individual's gender based on facial images. Earlier approaches mainly relied on traditional image processing and machine learning techniques, where features such as shape, texture, and color were extracted manually and then classified using statistical or learning-based models [25]. Among the most commonly adopted methods are Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), and Support Vector Machines (SVM), which have shown promising results in capturing discriminative facial attributes [12].

With the advancement of deep learning, recent studies have shifted towards Convolutional Neural Networks (CNNs) as they can automatically learn hierarchical features from raw image data through convolution, pooling, and non-linear activation layers [26]. Techniques such as transfer learning and fine-tuning of pre-trained models have been widely employed, especially when dealing with limited datasets. Models initially trained on large-scale datasets like ImageNet can be adapted to gender classification tasks with improved generalization [9]. Several CNN variants—including ResNet, GoogLeNet, VGGNet, MobileNet, and EfficientNet—have been extensively investigated for this purpose [14].

VGG-16, in particular, has been frequently adopted in gender recognition research due to its relatively simple yet effective architecture. It is composed of multiple convolutional layers with small receptive fields (3×3) grouped into blocks, each followed by max-pooling layers to reduce dimensionality while preserving important features [27]. Compared to more complex architectures such as ResNet or EfficientNet, VGG-16 offers a favorable balance between depth and computational cost, making it easier to implement and fine-tune while still achieving competitive accuracy in gender recognition tasks. Furthermore, its straightforward and uniform design facilitates the integration of additional modules, such as attention mechanisms, which is the focus of this study.

Previous research on CNN-based models has demonstrated consistent improvements in accuracy

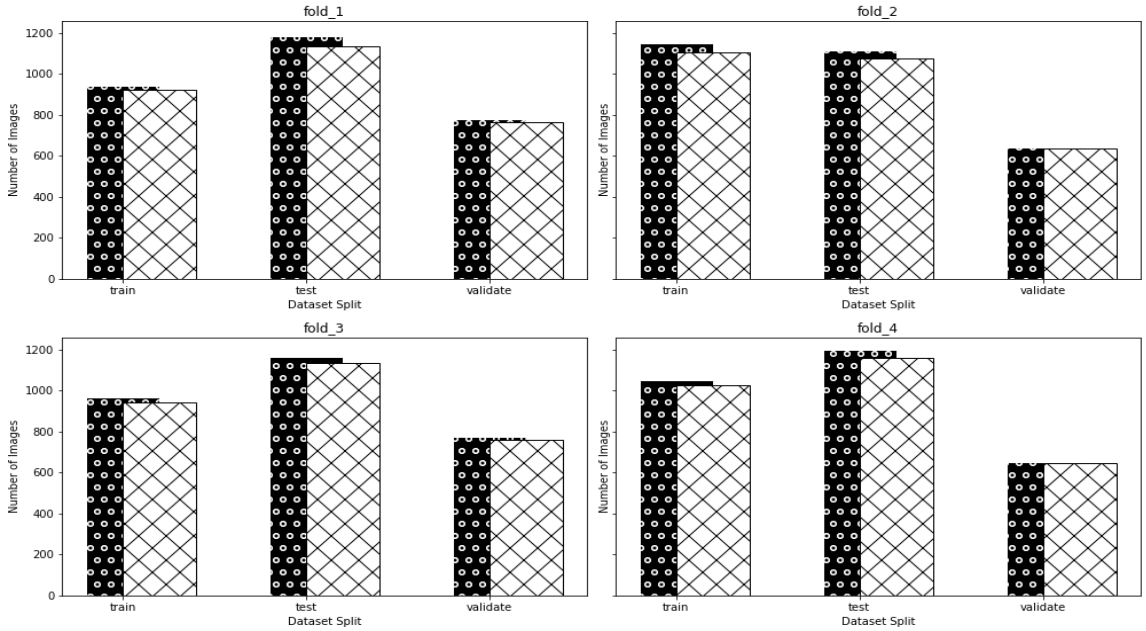

Figure 1. Gender distribution across training, validation, and test sets for each fold.

and efficiency for gender classification. ResNet introduces residual connections to address the vanishing gradient problem, GoogLeNet leverages inception modules with multi-scale filters, while MobileNet and EfficientNet emphasize lightweight architectures optimized for devices with limited resources [11], [12]. These advances highlight the ongoing need to explore architectural optimizations and hybrid approaches. Building on these insights, this study investigates the effectiveness of augmenting VGG-16 with attention mechanisms, specifically, Spatial Attention Blocks and Channel Attention Blocks to enhance the accuracy of face gender recognition.

3. Method

This study adopts a comparative modeling strategy to determine the most effective configuration of feature extraction and attention mechanisms for facial gender classification. The method involves the integration of attention modules into a deep convolutional neural network (CNN), followed by multi-scenario evaluation through fine-tuning and performance comparison. By employing an enhanced VGG-16 architecture, the objective is to boost the model's ability to learn and generalize gender-relevant facial features under varying real-world conditions. This section details the dataset, preprocessing techniques, model architecture, training scenarios, and evaluation metrics employed in the study.

4. Dataset

This research utilizes the Labeled Faces in the Wild Gender (LFW-Gender) [28] dataset, a curated subset of the widely used LFW dataset. The LFW-Gender dataset consists of 5,810 RGB facial images with a resolution of 200×200 pixels, which is balanced, containing 2,905 male and 2,905 female images. It is designed to reflect unconstrained settings, containing diverse variations in lighting, facial expressions, age, ethnicity, head poses, occlusions, and background conditions, which are representative of real-world environments.

To ensure robust and fair training, the dataset is partitioned into four non-overlapping folds, where each fold comprises distinct identities. This structure mitigates identity leakage across folds and ensures that no individual appears in both training and testing subsets within a fold, thus preventing data memorization. Table 1 presents the distribution of the dataset across the four folds.

Table 1. LFW gender dataset.

Fold Number	Train	Validation	Test
1	1886	1556	2368
2	2296	1284	2230
3	1932	1550	2328
4	2108	1302	2400
Total	**8222**	**5692**	**9326**

This fold-based setup supports cross-validation and allows a more comprehensive evaluation of model generalization. The dataset variability provides a challenging yet suitable

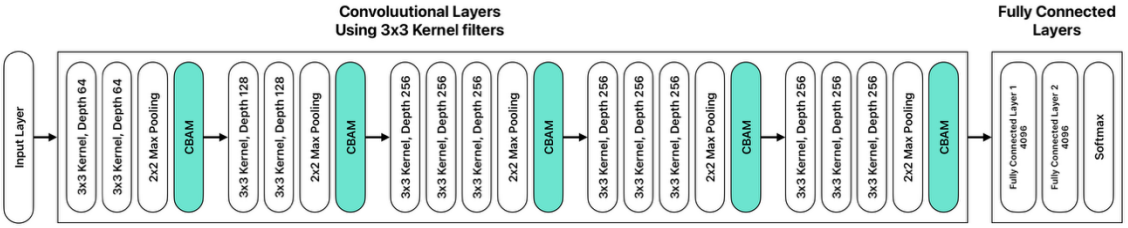

Figure 2. VGG-16 + CBAM architecture.

benchmark for evaluating the effectiveness of attention-augmented CNN models.

The LFW-Gender dataset was partitioned using a stratified approach. This preserved the male and female ratio within each fold across the training, validation, and testing sets. Figure 1 illustrates the distribution of male and female samples in all four folds. It shiwing that each split maintains a nearly 1:1 ratio with only minor deviations.

5. Preprocessing

Preprocessing is an essential phase in preparing raw input data for optimal training and inference performance. The preprocessing pipeline in this study consists of normalization and data augmentation.

5.1.1. Normalization

Normalization is implemented using Min-Max Scaling, which transforms pixel values to the [0, 1] range by dividing each pixel intensity by 255. This scaling standardizes the input across all images, helping to:

- Improve numerical stability during training.
- Accelerate convergence by keeping gradients within a manageable scale.
- Reduce the risk of vanishing/exploding gradients in deeper layers.

5.1.2. Data Augmentation

To counter overfitting and enhance the model's ability to generalize across unseen data, data augmentation is applied during training. The following augmentation techniques are randomly applied:

- Rotation: ±20° to simulate head tilts.
- Width and Height Shifts: ±10% to emulate slight facial displacement.
- Zoom: ±10% to mimic camera zoom variations.
- Horizontal Flipping: to introduce mirror-symmetric diversity in facial features

These transformations enrich the training data distribution and improve the model's invariance to small spatial distortions and orientation changes. The augmented data set leads to a more robust model capable of handling real-world inconsistencies.

6. Modeling

The core model used in this study is VGG-16, a CNN architecture known for its simplicity and strong performance in image classification. However, to address its limitations in handling salient feature localization, the Convolutional Block Attention Module (CBAM) is integrated. CBAM introduces spatial and channel-wise attention, enabling the model to focus on informative regions and feature maps.

6.1.1. Convolutional Block Attention Module

CBAM is composed of two sequential submodules:

- Channel Attention Block (CAB): Computes inter-channel dependencies by applying global average and max pooling across spatial dimensions, followed by a shared multi-layer perceptron (MLP). It enables the network to prioritize feature maps most relevant to gender traits (e.g., jawline curvature, skin texture).
- Spatial Attention Block (SAB): Applies max and average pooling along the channel axis and processes the result with a convolutional layer. It allows the model to localize key spatial regions like eyes, lips, and cheekbones.

The addition of CBAM is expected to strengthen the feature representation of VGG-16 by dynamically refining attention at each stage of convolution. The architecture is shown on Figure 2.

6.1.2. Training Strategy

The training follows a two-phase fine-tuning approach, allowing the model to retain general-purpose visual features while adapting to the

specific gender classification task:

- Phase 1 (Transfer Learning): All convolutional layers in VGG-16 are frozen; only the fully connected (dense) layers are trained. This phase adjusts the classifier to the domain-specific data while preserving pretrained visual features.
- Phase 2 (Fine-Tuning): Selected convolutional blocks are unfrozen (Blocks 4 and 5, or Blocks 2 to 5 in advanced scenarios) and trained alongside the dense layers. Fine-tuning enables the model to refine high-level feature representations critical for gender differentiation.

6.1.3. Experimental Scenario

To explore the impact of CBAM integration and fine-tuning depth, six modeling scenarios are tested, as summarized in Table 2.

Each scenario aims to balance model complexity, attention granularity, and generalization ability.

Table 2. Experimental Scenario

No	Scenario
1	- VGG-16 - The last convolutional blocks (block 4 and block 5) are unfrozen
2	- VGG-16 + CBAM block 4, block 5; all VGG-16 blocks are frozen - Block 4 and block 5 are unfrozen
3	- VGG-16 + CBAM block 4, block 5; all VGG-16 blocks are frozen - Block 4 and block 5 are unfrozen, and CBAM is added to block 5
4	- VGG-16 + CBAM block 4, block 5; all VGG-16 blocks are frozen - Block 4 and block 5 are unfrozen, and CBAM block is added
5	- VGG-16 + CBAM all blocks; VGG-16 blocks are frozen - Block 4 and block 5 are unfrozen
6	- VGG-16 + CBAM all blocks, and blocks 2 to 5 are unfrozen - Blocks 2 to 5 are unfrozen

6.1.4. Evaluation Metric

To comprehensively assess model performance, the following metrics are employed:

- Accuracy: Measures the proportion of correctly classified samples. It provides a direct indicator of overall model effectiveness.
- Precision: Represents the proportion of true positive predictions among all predicted positives. High precision minimizes false positives.
- Recall (Sensitivity): Indicates the proportion of actual positives correctly identified by the model. High recall reduces missed gender predictions.
- F1 Score: Harmonic mean of precision and recall. It balances both metrics, especially valuable when class distributions are close but not identical.

All metrics are computed using micro-averaging, which aggregates the contributions of all classes and ensures fair evaluation when dataset class distributions are approximately balanced. The use of multiple metrics provides a well-rounded view of the model's strengths and limitations, especially in complex gender prediction cases where false positives and negatives carry different implication.

7. Result and Discussion

Modelling was undertaken to enhance classification performance using the VGG-16 network, which was integrated with the Convolutional Block Attention Module (CBAM), varied across blocks 1 to 5. The implementation of CBAM aimed to incorporate an attention mechanism capable of accentuating crucial image features, thereby improving the model's proficiency in differentiating between 'female' and 'male' classes. A series of six scenarios were tested to evaluate different strategies for CBAM integration and layer fine-tuning.

The experimental scenarios progressively explored the impact of CBAM placement and layer unfreezing on VGG-16 model performance. Scenario 1 established a baseline by unfreezing the final convolutional blocks (Block 4 and 5) of a standard VGG-16 model [13], which showed the minimum accuracy was 82.13% and the maximum accuracy reached 88.72% across folds (Table 3). This baseline helped in understanding the limitations of the original architecture when applied to gender classification tasks using facial images. The relatively wide range of accuracy across folds also hinted at potential inconsistencies, motivating the need for improved feature extraction through attention mechanisms.

Subsequent scenarios introduced CBAM. Scenario 2, applying CBAM to Block 4 and 5 (unfrozen in step 2), generally improved performance, with the minimum accuracy was 86.00% and the maximum rached 90.82% (Table 4). This improvement demonstrates that the attention mechanism successfully emphasized discriminative features, particularly those found in deeper convolutional layers. It can be inferred that the incorporation of CBAM enhanced the model’s ability to focus on gender-relevant regions of the

face such as jawline, eyebrows, and cheekbone contours. However, the performance still varied across folds, suggesting that attention alone may not fully resolve model generalization.

Scenario 3 further refined the process by additionally applying CBAM to Block 5 during fine-tuning, resulting in more stable accuracies, with the minimum accruacy was 85.31% and the maximum accuracy reached 88.68% range (Table 5). The reduced variability compared to Scenario 2 indicated a more consistent learning pattern, though slightly lower peak performance. This suggests that aggressive tuning of deeper layers may stabilize training but not always yield peak accuracy, implying a trade-off between performance stability and maximum accuracy. Fine-tuning with selective attention placement appears to reduce overfitting but may also restrict the model's flexibility in learning complex spatial representations.

In Scenario 4, CBAM modules were integrated into both Block 4 and Block 5 while these blocks were unfrozen during fine-0tuning. It allowing the model to refine the high-level convolutional features and attention-based representation. This scenario produced the minimum accuracy was 84.91% and the maximum accuracy reached 90.87% (Table 6). The result show an improvement over Scenario 2 and 3, suggesting that introducing attention earlier in the higher-level feature hieararchy helps the model better highlight discriminative regions while still benefiting from fine-tuning.

Scenario 5 expanded CBAM application to all blocks (1–5) in the initial step followed by unfreezing Block 4 and 5. This also led to variable performance with minimum accuracy was 84.17% and the maximum accruacy reached 89.86% (Table 7), again hinting at overfitting in some folds. The inclusion of CBAM in the shallower layers may have unintentionally overemphasized low-level features such as lighting or texture, which are less informative for gender classification. This further underscores the necessity of selective CBAM placement to avoid enhancing irrelevant features.

Finally, Scenario 6 implemented CBAM across all blocks (1–5) with Blocks 2–5 unfrozen during initial training, followed by fine-tuning these unfrozen blocks. This comprehensive approach yielded the highest and most consistent accuracies across all folds, with the minimum accruacy was 90.04% and the maximum accuracy reached 91.78% (Table 8). These results indicate that full integration of attention mechanisms, when coupled with a well-structured fine-tuning process, can significantly improve both performance and stability. It appears that unfreezing Block 2, in addition to deeper layers, provided a better gradient flow and feature propagation through the network, enabling improved learning across multiple abstraction levels.

In all scenarios, Fold 3 again exhibited a noticeable performance drop, indicating that the reduced accuracy is likely attributable to the data characteristics of this fold rather than limitation in the model architecture.

In addition to accuracy, qualitative analysis of the misclassified images revealed that certain facial features such as hairstyle, makeup, and occlusion (e.g., glasses, face masks) often led to incorrect predictions. This aligns with previous findings where visual ambiguity poses a challenge for binary gender classifiers. Future work may benefit from integrating auxiliary classifiers or context-aware mechanisms (e.g., pose estimation or facial landmark detection) to handle such edge cases.

Overall, the experimental results at Table 9 validate the effectiveness of combining CBAM with VGG-16 for gender classification tasks. The progressive enhancement across scenarios highlights that attention mechanisms significantly contribute to model refinement, and their success heavily depends on thoughtful integration strategies. In Scenario 1, the baseline VGG-16 model with partially unfrozen convolutional blocks achieved a reasonable accuracy (≈85%), serving as a foundation for comparison.

Scenarios 2 and 3 introduced CBAM selectively to the deeper layers (block 4 and block 5), showing moderate improvement but with performance fluctuations across folds, suggesting that attention in deeper layers alone is insufficient for optimal feature learning. Scenario 4 demonstrated a slight gain by further expanding the integration of CBAM while still maintaining frozen earlier layers, showing the benefit of directing attention more consistently. Scenario 5 applied CBAM across all convolutional blocks, which helped capture richer spatial and channel dependencies but was still limited by freezing the earlier layers. Finally, Scenario 6, which not only incorporated CBAM across all layers but also unfroze blocks 2 to 5, achieved the highest and most stable performance (≈90.7%), indicating that fine-tuning a larger portion of the network alongside attention mechanisms enables the model to adapt more effectively to the dataset’s discriminative features.

These findings suggest that CBAM is most beneficial when it is integrated throughout the architecture and paired with selective fine-tuning of deeper convolutional blocks, rather than when applied only to isolated layers. The outcomes of this study can be further extended to other classification tasks within computer vision where fine-grained discrimination is essential, particularly in cases

where subtle feature differences are critical for accurate decision-making.

Table 3. Scenario 1 [13] result

Fold	Accuracy	Precision	Recall	F1-Score
1	85.20%	85.30%	85.20%	85.20%
2	**88.73%**	**88.88%**	**88.73%**	**88.71%**
3	82.17%	83.34%	82.17%	82.04%
4	85.01%	86.10%	85.01%	84.92%

Table 4. Scenario 2 result

Fold	Accuracy	Precision	Recall	F1-Score
1	87.36%	87.41%	87.36%	87.36%
2	**90.78%**	**90.83%**	**90.78%**	**90.77%**
3	86.00%	86.07%	86.00%	85.99%
4	86.79%	86.82%	86.79%	86.78%

Table 5. Scenario 3 result

Fold	Accuracy	Precision	Recall	F1-Score
1	86.71%	87.36%	86.71%	86.68%
2	**88.73%**	**88.75%**	**88.73%**	**88.72%**
3	85.35%	85.63%	85.35%	85.33%
4	85.60%	85.85%	85.60%	85.56%

Table 6. Scenario 4 result

Fold	Accuracy	Precision	Recall	F1-Score
1	87.75%	87.81%	87.75%	87.75%
2	**90.87%**	**90.90%**	**90.87%**	**90.87%**
3	84.91%	85.18%	84.91%	84.89%
4	86.79%	86.81%	86.79%	86.79%

Table 7. Scenario 5 result

Fold	Accuracy	Precision	Recall	F1-Score
1	87.14%	87.14%	87.14%	87.14%
2	**89.86%**	**89.86%**	**89.86%**	**89.86%**
3	84.17%	84.17%	84.17%	84.17%
4	86.23%	86.23%	86.23%	86.23%

Table 8. Scenario 6 result

Fold	Accuracy	Precision	Recall	F1-Score
1	90.81%	90.85%	90.81%	90.81%
2	**91.83%**	**91.87%**	**91.83%**	**91.83%**
3	90.17%	90.21%	90.17%	90.17%
4	90.09%	90.11%	90.09%	90.09%

Table 9. Summary of each scenario

Scenario	Accuracy	Precision	Recall	F1-Score
1	85.28%	85.90%	85.28%	85.22%
2	87.73%	87.78%	87.73%	87.72%
3	86.60%	86.90%	86.60%	86.57%
4	87.58%	87.68%	87.58%	87.58%
5	86.83%	86.98%	86.83%	86.82%
6	**90.73%**	**90.76%**	**90.73%**	**90.73%**

Table 10. Complexity measurement

Scenario	Parameters	GFLOPS
1	**15.2 B**	**1.45**
2	19.4 B	1.46
3	19.5 B	1.46
4	33.7 B	1.47
5	19.4 B	1.46
6	19.4 B	1.46

Figure 3 and Figure 4 present the confusion matrices for Scenario 1 Fold 2 and Scenario 6 Fold 2, respectively. In Scenario 1, the model correctly classified 918 female and 1,026 male instances, while 159 female and 88 male samples were misclassified. Scenario 6 demonstrates a noticeable improvement, with 1,048 correctly identified female and 1,104 male samples, and fewer misclassifications (113 female and 96 male). These results indicate that the model in Scenario 6 achieves higher discriminative performance and better class balance, suggesting enhanced generalization capability. The reduction in false classifications implies that the optimization or architectural adjustments applied in Scenario 6 contribute to more reliable gender recognition across both categories.

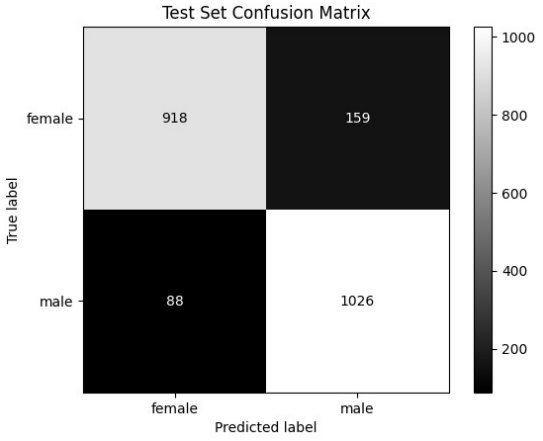

Figure 3. Confusion matrix of Scenario 1 Fold 2.

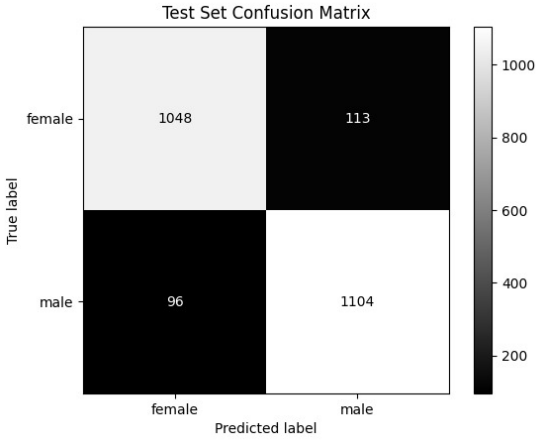

Figure 4. Confusion matrix of Scenario 6 Fold 2.

The complexity measurements in Table 10 show that adding CBAM to VGG-16 consistently increases the parameter count while only marginally affecting FLOPs. The baseline VGG-16 (Scenario 1) has 15.2B parameters and 1.45 GFLOPs. Configurations with CBAM applied to selected blocks (Scenarios 2, 3, and 5) expand the parameters to around 19.4–19.5B, with FLOPs

remaining at 1.46. Scenario 4 introduces the highest complexity, reaching 33.7B parameters and 1.47 GFLOPs, whereas Scenario 6 maintains the same complexity as Scenarios 2 and 5 despite unfreezing additional layers. These results suggest that the number of CBAM modules largely drives parameter growth, while the computational overhead in FLOPs remains relatively stable.

This indicates a clear trade-off, where the integration of CBAM increases model complexity primarily through parameter growth while offering potential improvements in feature representation with minimal additional computational cost.

8. Conclusion

This research successfully demonstrated the efficacy of integrating the Convolutional Block Attention Module (CBAM) across blocks 1 to 5 of the VGG-16 model to enhance gender classification performance on the LFW dataset. The inclusion of CBAM led to a notable increase in accuracy compared to the baseline VGG-16 model. The optimally adapted VGG-16 model, incorporating CBAM across all five blocks and subsequently fine-tuned (with particular attention to Blocks 4 and 5 in step 2 of the process), achieved a commendable accuracy of 90.35% on the validation set and 90.34% on the test set. Furthermore, precision, recall, and F1-Score metrics were consistently high, averaging around 90% for both 'female' and 'male' classifications. Although the fine-tuning and training regimen with CBAM demanded additional computational time, the significant improvement in accuracy underscores the module's effectiveness in this application.

A systematic exploration through various scenarios confirmed these benefits. While a baseline VGG-16 (Scenario 1) with unfrozen later blocks showed potential, its performance varied across folds. Introducing CBAM to Block 4 and 5 (Scenario 2) yielded higher accuracies, reaching up to 90.82% in one fold. Further refinements, such as adding CBAM specifically to Block 5 during fine-tuning (Scenario 3), contributed to more stable accuracy across different data folds. Scenarios 4 and 5 also exhibited strong accuracies, albeit with some inter-fold fluctuations. Collectively, these experiments consistently highlighted that integrating CBAM substantially improved the VGG-16 model's gender classification capabilities. Based on the findings, several recommendations for future research and development are proposed:

- Employ optimization algorithms to systematically explore a wider range of fine-tuning hyperparameters, including learning rate, batch size, and the number of epochs, to identify an even more optimal model configuration.
- Utilize more diverse facial datasets to enhance the model's generalization capabilities across a broader spectrum of variations in lighting, pose, and expressions.
- Incorporate more sophisticated data augmentation techniques, such as color jittering, motion blur, or advanced geometric distortions. These methods can further improve the model's robustness and its ability to perform reliably under diverse real-world conditions.
- Implement automatic face detection and cropping mechanisms (e.g., using MTCNN or Haar Cascades) prior to classification. This would ensure the model focuses exclusively on relevant facial features, potentially reducing background noise and further improving recognition accuracy by concentrating learning on the most pertinent image regions.

References

[1] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, “Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data,” in *Proceedings of the Eighteenth International Conference on Machine Learning (ICML ’01)*, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001, pp. 282–289.

[2] P. Krähenbühl and V. Koltun, “Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials,” *Advances in Neural Information Processing Systems 24 (NIPS 2011)*, pp. 109–117, Oct. 2012, [Online]. Available: http://arxiv.org/abs/1210.5644

[3] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, “CNN Features off-the-shelf: An Astounding Baseline for Recognition,” *CoRR*, vol. abs/1403.6382, 2014, [Online]. Available: http://arxiv.org/abs/1403.6382

[4] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, “Return of the Devil in the Details: Delving Deep into Convolutional Nets,” *Proceedings of the British Machine Vision Conference (BMVC)*, Nov. 2014, [Online]. Available: http://arxiv.org/abs/1405.3531

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” *Commun ACM*, vol. 60, no. 6, pp. 84–90, May 2017, doi: 10.1145/3065386.

[6] A. Vezhnevets, V. Ferrari, and J. M. Buhmann, “Weakly supervised semantic segmentation with a multi-image model,” in *2011 International Conference on Computer Vision*, IEEE, Nov. 2011, pp. 643–650. doi: 10.1109/ICCV.2011.6126299.

[7] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, “The Pascal Visual Object Classes (VOC) Challenge,” *Int J Comput Vis*, vol. 88, no. 2, pp. 303–338, Jun. 2010, doi: 10.1007/s11263-009-0275-4.

[8] A. Dhomne, R. Kumar, and V. Bhan, “Gender Recognition Through Face Using Deep Learning,” *Procedia Comput Sci*, vol. 132, pp. 2–10, 2018, doi:

10.1016/j.procs.2018.05.053.

[9] W.-S. Chen and R.-H. Jeng, "A new patch-based LBP with adaptive weights for gender classification of human face," *Journal of the Chinese Institute of Engineers*, vol. 43, no. 5, pp. 451–457, Jul. 2020, doi: 10.1080/02533839.2020.1751724.

[10] Mohammed Jawad Al_Dujaili, H. TH. S. ALRikabi, Nisreen Khalil abed, and Ibtihal Razaq Niama ALRubeei, "Gender Recognition of Human from Face Images Using Multi-Class Support Vector Machine (SVM) Classifiers," *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 17, no. 08, pp. 113–134, Apr. 2023, doi: 10.3991/ijim.v17i08.39163.

[11] Z. Zhou, F. Liu, and Q. Wang, "R-Transformer Network Based on Position and Self-Attention Mechanism for Aspect-Level Sentiment Classification," *IEEE Access*, vol. 7, pp. 127754–127764, 2019, doi: 10.1109/ACCESS.2019.2938854.

[12] Y. Sun, B. Xue, M. Zhang, and G. G. Yen, "Completely Automated CNN Architecture Design Based on Blocks," *IEEE Trans Neural Netw Learn Syst*, vol. 31, no. 4, pp. 1242–1254, Apr. 2020, doi: 10.1109/TNNLS.2019.2919608.

[13] S. Mittal and S. Mittal, "Gender Recognition from Facial Images using Convolutional Neural Network," in *2019 Fifth International Conference on Image Information Processing (ICIIP)*, IEEE, Nov. 2019, pp. 347–352. doi: 10.1109/ICIIP47207.2019.8985914.

[14] A. Greco, A. Saggese, M. Vento, and V. Vigilante, "A Convolutional Neural Network for Gender Recognition Optimizing the Accuracy/Speed Tradeoff," *IEEE Access*, vol. 8, pp. 130771–130781, 2020, doi: 10.1109/ACCESS.2020.3008793.

[15] V. Albiero, K. Zhang, and K. W. Bowyer, "How Does Gender Balance In Training Data Affect Face Recognition Accuracy?," in *2020 IEEE International Joint Conference on Biometrics (IJCB)*, IEEE, Sep. 2020, pp. 1–10. doi: 10.1109/IJCB48548.2020.9304924.

[16] A. Althnian, N. Aloboud, N. Alkharashi, F. Alduwaish, M. Alrshoud, and H. Kurdi, "Face Gender Recognition in the Wild: An Extensive Performance Comparison of Deep-Learned, Hand-Crafted, and Fused Features with Deep and Traditional Models," *Applied Sciences*, vol. 11, no. 1, p. 89, Dec. 2020, doi: 10.3390/app11010089.

[17] C. Jia, X. Li, R. Qian, and H. Sun, "Facial Expression Recognition Based on Pruning Optimization Technology," *Highlights in Science, Engineering and Technology*, vol. 41, pp. 101–110, Mar. 2023, doi: 10.54097/hset.v41i.6784.

[18] V. Carletti, A. Greco, A. Saggese, and M. Vento, "Attention-based Gender Recognition on Masked Faces," in *Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, SCITEPRESS - Science and Technology Publications, 2022, pp. 672–678. doi: 10.5220/0010978700003124.

[19] H. Ling, J. Wu, J. Huang, J. Chen, and P. Li, "Attention-based convolutional neural network for deep face recognition," *Multimed Tools Appl*, vol. 79, no. 9–10, pp. 5595–5616, Mar. 2020, doi: 10.1007/s11042-019-08422-2.

[20] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional Block Attention Module," *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 3–19, Jul. 2018, [Online]. Available: http://arxiv.org/abs/1807.06521

[21] S. Poornima, N. Sripriya, S. Preethi, and S. Harish, "Classification of Gender from Face Images and Voice," in *Intelligence in Big Data Technologies—Beyond the Hype: Proceedings of ICBDCC 2019*, 2021, pp. 115–124. doi: 10.1007/978-981-15-5285-4_11.

[22] M. Zhu, L. Jiao, F. Liu, S. Yang, and J. Wang, "Residual Spectral–Spatial Attention Network for Hyperspectral Image Classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 1, pp. 449–462, Jan. 2021, doi: 10.1109/TGRS.2020.2994057.

[23] K. Takahashi, K. Yamamoto, A. Kuchiba, and T. Koyama, "Confidence interval for micro-averaged F1 and macro-averaged F1 scores," *Applied Intelligence*, vol. 52, no. 5, pp. 4961–4972, Mar. 2022, doi: 10.1007/s10489-021-02635-5.

[24] K. Takahashi, K. Yamamoto, A. Kuchiba, and T. Koyama, "Confidence interval for micro-averaged F1 and macro-averaged F1 scores," *Applied Intelligence*, vol. 52, no. 5, pp. 4961–4972, Mar. 2022, doi: 10.1007/s10489-021-02635-5.

[25] M. Afifi, "11K Hands: Gender recognition and biometric identification using a large dataset of hand images," *Multimed Tools Appl*, vol. 78, pp. 20835–20854, Sep. 2018, [Online]. Available: http://arxiv.org/abs/1711.04322

[26] F. Waris, F. Da, and S. Liu, "Deep learning based features extraction for facial gender classification using ensemble of machine learning technique," *Multimed Syst*, vol. 30, no. 4, p. 200, Aug. 2024, doi: 10.1007/s00530-024-01399-5.

[27] S. Tammina, "Transfer learning using VGG-16 with Deep Convolutional Neural Network for Classifying Images," *International Journal of Scientific and Research Publications (IJSRP)*, vol. 9, no. 10, p. p9420, Oct. 2019, doi: 10.29322/IJSRP.9.10.2019.p9420.

[28] A. Jalal and U. Tariq, "The LFW-Gender Dataset," in *Asian Conference on Computer Vision*, 2017, pp. 531–540. doi: 10.1007/978-3-319-54526-4_39.