AUT Journal of Modeling and Simulation

AUT J. Model. Simul., 58(2) (2026) 189-204
DOI: 10.22060/miscj.2026.25626.5483

MBFANet: A Lightweight Network with Multi-Branch Fusion and Attention Modules for Real-time Segmentation

Zainab Rawan Abdulraheem [1], Mansoor Fateh [1,*], Saedeh Ferdowsi [2], Sadjad Rezvani [1]

[1] Faculty of Computer Engineering, Shahrood University of Technology, Shahrood, Iran.
[2] School of Mathematics, Statistics and Actuarial Science, University of Essex, Colchester, England.

ABSTRACT: Masked face segmentation is crucial for face recognition, security, and medical applications, yet it remains challenging due to occlusions, varying lighting conditions, and diverse mask patterns. To address these issues, we propose MBFANet, a lightweight and efficient deep learning model that integrates multi-branch feature fusion and attention mechanisms for real-time segmentation. Our architecture builds on the efficient convolutional backbone RepVGG and adds two key modules: Channel Aggregation and Coordinate Refinement (CACR) to improve spatial and channel-wise feature representation, and the Atrous Pyramid Attention Module (APAM) to capture multi-scale contextual information. Additionally, a Multi-Branch Feature Fusion (MBFF) module is employed in the decoder to refine segmentation outputs. To optimize segmentation performance, we introduce a hybrid loss function that combines Dice Loss and Focal Loss, effectively addressing class imbalance while improving boundary precision. Experimental evaluations on the Masked Face Segmentation Dataset (MFSD) demonstrate that MBFANet achieves the highest IoU, F1-score, and recall among the compared models while running at the fastest inference speed (153 FPS). Although a few lightweight baselines use fewer parameters, MBFANet offers the most favorable overall balance of accuracy, speed, and model size (10M parameters), making it well-suited for real-time masked face segmentation. Ablation studies confirm the effectiveness of the proposed modules, showing consistent improvements across all evaluation metrics. With its balance between accuracy, efficiency, and computational cost, MBFANet presents a highly effective solution for real-time masked face segmentation, making it well-suited for practical deployment in security, biometric authentication, and mobile applications.

Review History:
Received: Mar. 15, 2026
Revised: Jun. 21, 2026
Accepted: Jul. 01, 2026
Available Online: Sep. 15, 2026

Keywords:
Masked Face Segmentation
Deep Learning
Multi-Branch Fusion
Attention Mechanisms
Real-Time Segmentation

1- Introduction

The rise of deep convolutional neural networks (CNNs) has significantly advanced face recognition, making it essential for applications like surveillance [1], border control [2], education [3], and healthcare [4]. However, real-world conditions introduce several challenges, including changes in lighting [5], low image resolution [6], varying facial poses [7], dynamic expressions [8]and occlusions from masks, glasses, or scarves [9, 10]. To improve accuracy, recent trends in computer vision have focused on building deeper and more complex CNNs with hundreds of layers and thousands of feature channels. While these models achieve state-of-the-art results, they come at the cost of high computational demand. This makes them impractical for real-time applications, especially on mobile devices, embedded systems, and edge computing platforms. In time-sensitive tasks like biometric authentication or augmented reality, heavy models struggle with efficiency due to their large memory footprint and slow processing speed [11-13]. To address these challenges, lightweight CNN architectures have gained attention as a promising solution. These models aim to maintain high segmentation and recognition accuracy while significantly reducing computational costs, enabling efficient deployment in real-world applications.

Occlusion in face recognition happens when parts of the face are blocked by objects, accessories, or other elements. This obstruction significantly impacts recognition accuracy because important facial features may be hidden. Common causes of occlusion include sunglasses, hats, scarves, and even environmental factors such as shadows or poor lighting. When key features are covered, face recognition algorithms struggle to extract meaningful information. As a result, the system's ability to correctly identify individuals is reduced, leading to higher error rates and misidentifications.

One of the most challenging forms of occlusion in recent years has been the widespread use of face masks, especially during the COVID-19 pandemic. While masks have played a vital role in public health, they have also posed serious challenges for biometric identification systems. By covering a large portion of the lower face, masks hide features like the

*Corresponding author's email: mansoor_fateh@shahroodut.ac.ir

Copyrights for this article are retained by the author(s) with publishing rights granted to Amirkabir University Press. The content of this article is subject to the terms and conditions of the Creative Commons Attribution 4.0 International (CC-BY-NC 4.0) License. For more information, please visit https://www.creativecommons.org/licenses/by-nc/4.0/legalcode.

UNCORRECTED PROOF

nose shape, mouth structure, and the distances between these features. Many face recognition models rely on these regions for accurate identification. Studies have shown that masks can obscure nearly half of the biometric information used in recognition, making it difficult for traditional algorithms to function effectively [14].

Beyond recognition accuracy, mask-induced occlusion has also created security concerns. Fraudsters and criminals have exploited masks to evade detection in surveillance systems, making it harder for law enforcement to track individuals. This issue has highlighted the need for improved face recognition models that can handle occlusions more effectively. Addressing this challenge requires new approaches that can accurately segment masked faces while maintaining real-time processing capabilities.

As illustrated in Figure 1, the masked face recognition process consists of several essential stages designed to ensure accurate identification despite occlusions. The first step is masked face detection, where the system identifies whether a face in an image is partially covered. After detection, preprocessing methods like normalization, alignment, and illumination correction are performed to augment feature consistency as well as enhance recognition performance.

Two primary preprocessing methods have been introduced following face detection: representation and reconstruction. In the representation-based approach, the face mask is first segmented out and excluded from the feature extraction process. Face recognition is then performed on the unmasked portion of the face [15]. On the contrary, reconstruction techniques address occlusion challenges in the image by attempting to restore the obscured facial areas, aiming to produce an image closer to the original face. This recovery is heavily reliant on binary mask segmentation, a key point highlighted in prior research

[16-18]. Importantly, reconstruction methods have shown superior performance, particularly when recognizing faces with significant occlusions, such as those caused by masks [19].

A significant portion of the techniques for recognizing occluded faces relies heavily on mask segmentation techniques. Through precise segmentation of the mask, these techniques try to reduce the adverse effect of occlusion on the facial features, thereby enhancing the performance of face recognition. Mask segmentation allows the system to focus on the visible portions of the face, thereby enhancing recognition performance even when significant portions of the face are covered. This strategy has proven to be effective in overcoming challenges posed by mask-wearing, as demonstrated in recent studies [20].

Feature extraction focuses on obtaining discriminative representations from facial images using deep learning-based or traditional techniques, ensuring that key identity-related features are preserved despite occlusions. These extracted features are then utilized in the classification stage, where they are compared against stored templates to determine the most likely identity. Finally, post-processing and verification

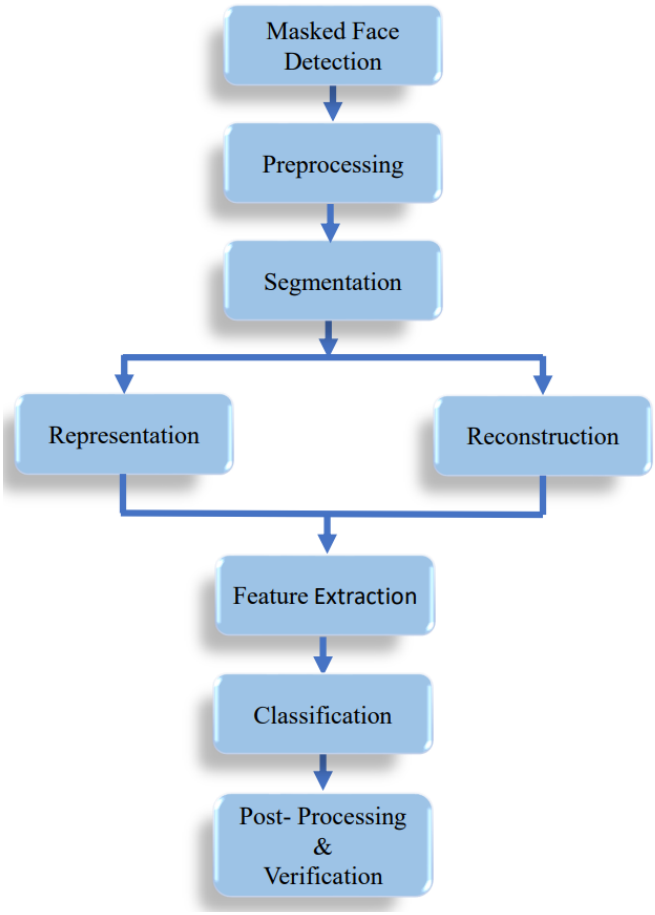


Fig. 1. Overview of the masked face recognition pipeline.

refine the recognition results through confidence thresholding, fusion strategies, or additional checks, improving the reliability and accuracy of the system.

Mask segmentation presents significant challenges due to the wide variety of mask types, colors, and designs worn by individuals, making it difficult to generalize across different cases.

Despite recent progress in lightweight segmentation, existing designs exhibit several limitations when applied to masked face segmentation. First, most channel-attention modules recalibrate features along the channel dimension alone and disregard the positional structure of the feature map, which limits their ability to localize mask boundaries precisely. Second, conventional multi-scale modules either enlarge the receptive field at the cost of additional parameters or aggregate multi-scale responses without an attention mechanism to emphasize the most informative scales. Third, common feature-fusion strategies based on simple addition or concatenation treat the contributing branches independently and do not explicitly model the interaction between features extracted at different levels. The three modules proposed in

UNCORRECTED PROOF

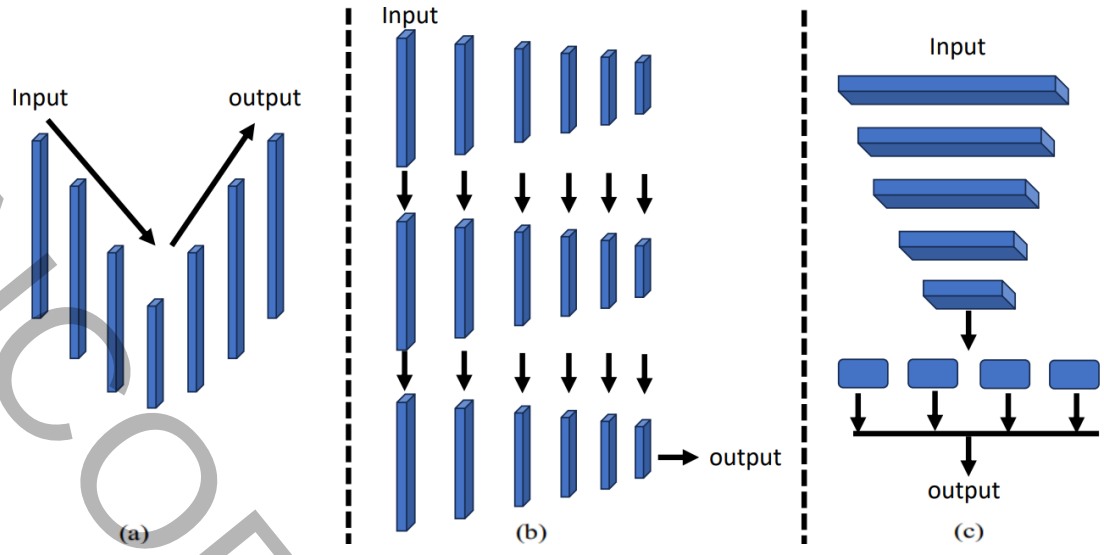


Fig. 2. Comparison of U-Shape, multi-branch, and pyramid structures. Two-branch structures. From left to right: (a) U-shape; (b) multi-branch; (c) Pyramid.

UNCORRECTED PROOF

this work are designed to address these limitations directly.

Our three main contributions are as follows:

1. We propose the **Channel Aggregation and Coordinate Refinement (CACR)** module, which jointly models channel-wise and coordinate-aware spatial information. By aggregating positional cues along the horizontal and vertical directions, CACR overcomes the limited spatial localization of conventional channel-only attention and improves boundary delineation.

2. We introduce the **Atrous Pyramid Attention Module (APAM)**, which combines dilated convolutions at multiple rates with an attention mechanism. This enables the network to capture multi-scale context with negligible additional cost while adaptively emphasizing the most relevant scales, addressing the lack of attention-guided scale selection in prior multi-scale modules.
3. We design a **Multi-Branch Feature Fusion (MBFF)** module that integrates features from different encoder stages through an interaction-aware fusion scheme. Unlike simple additive or concatenative fusion, MBFF explicitly models inter-branch dependencies, yielding more expressive multi-scale representations for accurate segmentation.

The structure of this paper is as follows: Section 2 provides a review of recent related research undertaken by various researchers; Section 3 explains the model architecture; Section 4 presents an experiment to demonstrate the effectiveness of the proposed approach; Section 5 concludes the paper and states the future work.

2- Related work

Based on our proposed method, we summarize relevant works in the following key aspects: real-time semantic segmentation models, which explore efficient architectures for fast and accurate segmentation; feature fusion strategies, which enhance representation learning by effectively integrating multi-scale and contextual information; dilated convolution, which expands the receptive field without increasing computational cost; and attention mechanisms, which improve feature selection and refinement for better segmentation performance.

2- 1- Real-time segmentation networks

The growing complexity of model architectures in semantic segmentation tends to put a huge computational pressure on hardware, and real-time processing becomes a very difficult proposition [21]. In response to this issue, researchers have concentrated their efforts on developing and refining lightweight network architectures that achieve a trade-off between accuracy and efficiency.

Over the years, various structural designs have been proposed, all aimed at improving the speed and performance of semantic segmentation models. Figure 2 summarizes network architectures used in segmentation. Among them, three principal architectures have emerged as the most important: U-shaped networks that utilize encoder-decoder models for efficient feature extraction and reconstruction; multi-branch networks that introduce parallel processing streams to identify a variety of feature representations; and pyramid networks that enhance contextual perception by adding multi-scale information. These innovative designs have played a major role in advancing real-time semantic

segmentation, making it possible to achieve fast and accurate scene comprehension even on resource-limited devices.

An example of a real-time U-Net-based segmentation model is Mobile-Unet [22], which was introduced for efficient hair segmentation. Mobile-Unet integrates the U-Net architecture with MobileNetV2 optimization techniques, significantly enhancing both speed and accuracy. EFFB7-UNet [23]EFFB7-UNet is a state-of-the-art semantic segmentation framework designed for real-time applications in Indoor Autonomous Vision Systems (IAVSs). Built on the U-Net architecture, EFFB7-UNet leverages EfficientNetB4 as its encoder to optimize feature extraction and incorporates a spatial and channel Squeeze-and-Excitation (scSE) attention block to focus on critical regions and features.

Bi-Unet [24] is a dual-stream lightweight network designed to tackle the challenges of real-time highway surface segmentation. As a multi-branch network, it incorporates a novel Lightweight Segmentation Module (LSM), which reduces model complexity while maintaining high accuracy. Furthermore, the network includes a Road Attention Network (RAN) module, which improves segmentation reliability in complex scenarios by emphasizing critical road features. AM-MulFSNet [25] proposes a real-time semantic segmentation algorithm designed to balance accuracy and computational efficiency. The model is built on the Fast Convolutional Neural Network (Fast-SCNN) framework and incorporates an attention mechanism alongside multi-branch feature fusion. To enhance spatial detail features, a spatial detail feature enhancement branch is introduced.

As an example of a real-time pyramid approach, the paper [26] introduces a novel pyramid structure. By stacking dilated blocks, the authors develop a dedicated backbone specifically designed for semantic segmentation, eliminating the need for additional modules. LightCM-PNet [27] proposes a lightweight and accurate segmentation network designed to address the challenges of prostate segmentation. A pyramid structure is employed for feature encoding and decoding, effectively reducing the number of parameters and computational complexity while utilizing global context information. To enhance segmentation accuracy, the authors introduce an interactive hybrid attention module that sequentially generates attention maps along the channel and spatial dimensions, focusing on the prostate region in the output features.

2- 2- Feature fusion strategies

Feature fusion is key to improving the performance of semantic segmentation models by effectively integrating multi-scale, contextual, and hierarchical information. Different layers in a neural network capture various levels of abstraction, ranging from low-level spatial details to high-level semantic features; efficient fusion strategies are important for preserving valuable information and enhancing segmentation accuracy.

An example of feature fusion in semantic segmentation is PL-UNet [28], a model designed for real-time power line segmentation using UAV imagery. A multi-scale attention gate (MSAG) is integrated into the decoder, refining key feature extraction with minimal computational overhead. Additionally, the dynamic weighted fusion (DWF) module adaptively balances feature integration, enhancing flexibility in feature representation.

Another effective approach to feature fusion in real-time semantic segmentation is LFFNet (Lightweight Feature-Enhanced Fusion Network) [29]. In the encoder, a multi-dilation rate fusion module expands the receptive field while preserving essential local information, effectively handling variations in target sizes and improving feature extraction.

2- 3- Dilated Convolutions

Atrous convolution, or dilated convolution, is an effective technique employed for semantic segmentation that yields an expanded receptive field without any additional computation cost. Unlike standard convolutions with a fixed receptive field, dilated convolutions insert holes—dilation rates—between kernel elements that allow the network to aggregate multi-scale contextual information without altering the spatial resolution.

A recent advancement in dilated convolution is Frequency-Adaptive Dilated Convolution (FADC) [30]which improves traditional dilated convolution by dynamically adjusting dilation rates based on local frequency components. FADC adapts the receptive field spatially, allowing the network to capture both fine-grained details and broader contextual information more effectively.

Another innovative approach that leverages dilated convolution for efficient image processing is Attentive Dilated U-Net (ADU-Net) [31]which is designed specifically for low-light image enhancement. ADU-Net addresses this issue through a lightweight U-shaped architecture that integrates dilated residual modules for feature extraction and attentive fusion modules for effective feature aggregation. The dilated residual modules enhance the network's ability to extract representative features, while the attentive fusion modules combine low-level texture details with high-level semantic information, ensuring a more balanced and informative representation.

2- 4- Attention mechanisms

Attention mechanisms have become a fundamental component in deep learning models for semantic segmentation, serving as a key driver for feature selection and information aggregation. Attention modules dynamically balance different regions of an image, enabling the network to focus on the most informative and relevant features, leading to more precise and robust segmentation.

Different kinds of attention mechanisms are commonly utilized in segmentation tasks:

- **Channel Attention**: It reinforces the most discriminative channel features by remapping channel-wise feature responses. Spatial attention emphasizes significant regions in an image, thereby enhancing the foreground-background discrimination capability of the network.
- **Spatial attention**: emphasizes significant regions in an

UNCORRECTED PROOF

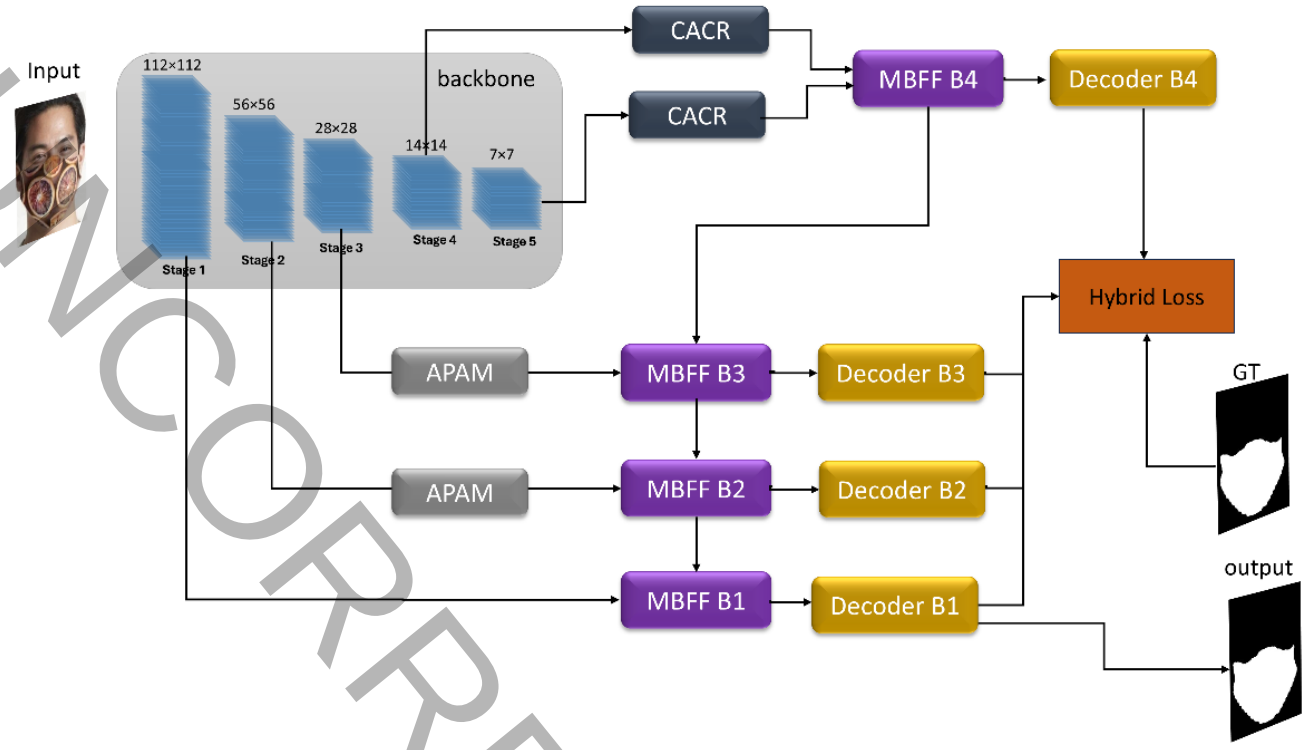


Fig. 3. Architecture of our proposed MBFANet network.

image, thereby enhancing the foreground-background discrimination capability of the network.

- **Self-Attention (Transformers)**: Maintains long-range dependencies by computing relations between all the pixels in an image, providing a more global scene structure comprehension.
- **Hybrid Attention**: It blends various attention types, i.e., channel and spatial, to create a more integrated feature refinement process.

The authors propose ALANet [32]a lightweight network designed to balance segmentation accuracy and inference speed for real-time applications. It features a channel-wise attention-based depth-wise asymmetric block (CADAB) in the encoder to efficiently capture key features with fewer parameters. The decoder uses two attention mechanisms: a spatial attention-based pyramid pooling (SAPP) module for multi-scale context aggregation and a pixel-wise attention-based multi-scale feature fusion (PAMFF) module for fusing features and enhancing image restoration.

CAWANet [33] is a lightweight and efficient method for real-time surface defect segmentation in manufacturing industries. To address challenges such as low contrast between defects and backgrounds and the need for real-time detection, context-aware adaptive weighted convolution (CAWAConv) is introduced. This technique extracts deep features while minimizing noise interference, allowing the model to remain computationally efficient without sacrificing defect recognition.

3- Methodology

3- 1- Overall architecture

Figure 3 shows the encoder-decoder architecture of the proposed MBFANet. In the encoder part of our proposed method, we employ RepVGG [34]a reparameterizable CNN architecture that enhances efficiency while maintaining strong feature extraction capabilities. Unlike traditional deep networks that use residual structures to improve performance, RepVGG adopts a simple VGG-style stack of 3×3 convolutional layers during inference. To compensate for the lightweight nature of RepVGG and enhance the extraction of deeper semantic features, we integrate the Channel Aggregation and Coordinate Refinement (CACR) module. This module refines the spatial and channel-wise feature representations, improving the model's ability to capture fine-grained details.

Additionally, we introduce the Atrous Pyramid Attention Module (APAM) to capture multi-scale contextual information by employing dilated convolutions with varying rates (e.g., 1, 6, and 12). The APAM leverages global average pooling and an attention mechanism to enhance feature responsiveness across different receptive fields. This ensures that both local and global contextual dependencies are effectively represented in the feature maps.

In the decoder stage, we employ a Multi-Branch Feature Fusion (MBFF) module to aggregate and refine multi-scale features extracted from different stages of RepVGG. This fusion mechanism strengthens feature representation,

UNCORRECTED PROOF

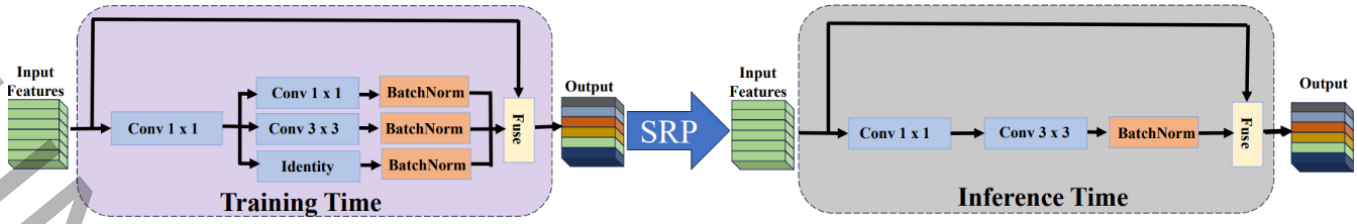


Fig. 4. RepVGG structure during training and inference. The three-branch design used in training is merged into a single 3×3 convolution after structural reparameterization (SRP) for efficient inference.

ultimately improving the segmentation accuracy of the network.

3- 2- Encoder: RepVGG

The encoder of our proposed framework is built upon RepVGG, a structurally reparameterizable convolutional neural network designed to achieve high performance with a simple architecture. RepVGG follows a five-stage design, where each stage consists of multiple stacked 3×3 convolutional layers with increasing channel dimensions. The first stage begins with 64 channels, and subsequent stages

The number of channels is progressively increased to 128, 256, 512, and 1024, following a VGG-style hierarchy. Downsampling is achieved using stride-2 convolutions, eliminating the need for pooling layers.

Illustrated in Figure 4, during training, RepVGG incorporates identity shortcuts and 1×1 convolutions alongside the 3×3 convolutional layers in each stage, forming a multi-branch topology that enhances gradient flow and feature learning. However, at inference time, these additional branches are mathematically fused into a single equivalent 3×3 convolution per layer, resulting in a plain, feedforward network that significantly reduces computational overhead. This transformation enables RepVGG to achieve ResNet-level accuracy while maintaining the hardware efficiency of VGG-style networks, making it highly suitable for real-time applications.

3- 3- Channel Aggregation and Coordinate Refinement (CACR)

The Channel Aggregation and Coordinate Refinement (CACR) module is designed to enhance feature representations by incorporating both spatial and channel-wise attention mechanisms. Unlike conventional attention modules that focus solely on channel dependencies, CACR simultaneously captures global spatial relationships to refine feature maps effectively.

As illustrated in Figure 5, the CACR module first reduces the channel dimensions of the input feature map using a 1×1 convolution, ensuring computational efficiency. Next, global average pooling (GAP) is applied along both the height (H) and width (W) dimensions, generating two distinct attention maps with shapes C × H × 1 and C × 1 × W, respectively. These maps are then processed through a 1×1 convolution, Batch Normalization (BN), and an activation function, yielding refined attention weights along two orthogonal spatial directions.

To enhance feature representation, the obtained attention weight maps are multiplied element-wise with the input feature map, selectively emphasizing important spatial locations. Finally, a residual connection is employed, ensuring the refined output retains essential contextual information. By incorporating both channel aggregation and coordinate refinement, CACR improves the network's ability to focus on critical spatial details while maintaining computational efficiency.

This approach enhances the discriminative power of high-level features, ensuring that both local and global spatial dependencies are preserved, leading to more precise feature extraction for segmentation tasks.

3- 4- Atrous Pyramid Attention Module (APAM)

The Atrous Pyramid Attention Module (APAM) is designed to preserve the hierarchical information and enhance the network's ability to distinguish fine-grained masked facial features. As masked face segmentation involves both local texture details and global structural patterns, APAM integrates dilated convolutions and an attention mechanism to improve feature representation across different scales.

As illustrated in Figure 6, APAM employs dilated convolutions with varying dilation rates (1, 6, and 12) to expand the receptive field without increasing the number of parameters. This enables the network to extract both local fine details and broader contextual features, which are important for varying mask

patterns. The extracted multi-scale features are then concatenated and passed through a 1×1 convolution, followed by a GAP operation. The global-pooling branch yields an image-level descriptor that is tiled back to the spatial resolution of the dilated outputs before concatenation, mirroring the global-context branch in ASPP/DeepLabV3+.

To further enhance feature selection, an attention mechanism is applied, generating an attention feature map that selectively emphasizes relevant regions while suppressing

UNCORRECTED PROOF

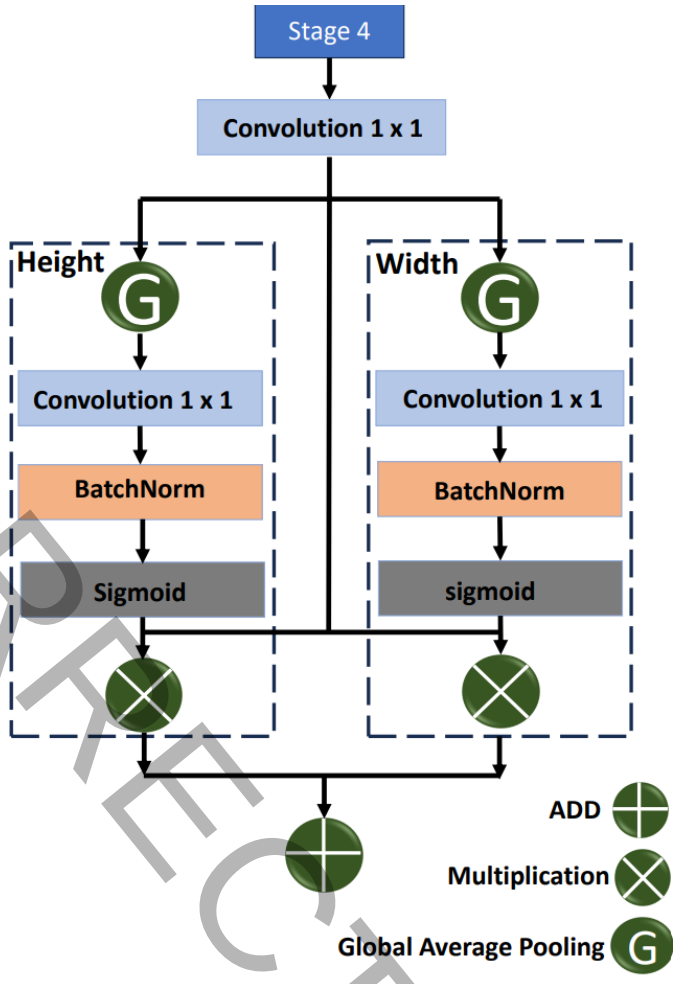


Fig. 5. Architecture of the CACR Module: Channel Aggregation and Coordinate Refinement Process.

less informative areas. This attention map is element-wise multiplied with the original feature representation, ensuring that the model focuses on critical spatial details essential for accurate mask segmentation. The process is shown in Algorithm 1.

By integrating APAM, our model effectively balances local feature refinement and global contextual understanding, making it highly suitable for precise and robust masked face segmentation.

3- 5- Multi-Branch Feature Fusion (MBFF)

MBFF module plays a critical role in the decoder stage by aggregating and refining multi-scale feature representations extracted from different stages of the encoder. Figure 7 illustrates the detailed structure of the MBFF module, which is designed to improve feature expressiveness and segmentation accuracy through an effective hierarchical fusion strategy.

The MBFF module consists of multiple parallel branches, each responsible for processing feature maps from different levels of the encoder. Each branch first applies a 1×1 convolution to reduce the channel dimensions while maintaining spatial information. This is followed by Batch Normalization (BatchNorm)

and ReLU activation, ensuring stable training and improved feature non-linearity. Once processed, the features from multiple branches are fused using element-wise multiplication, which enhances inter-feature

interactions while preserving critical spatial details. Additionally, reshaping operations are applied to align feature dimensions before fusion, ensuring smooth integration of multi-scale information.

The output from MBFF at each level is progressively refined by combining lower-level details with higher-level semantic features. This hierarchical refinement allows the network to maintain fine-grained

localization accuracy while leveraging the semantic richness of deep features. The parallel processing branches ensure that features from different resolutions are effectively utilized, capturing both high-level semantic context and fine details. By employing channel-wise and spatial-wise feature

UNCORRECTED PROOF

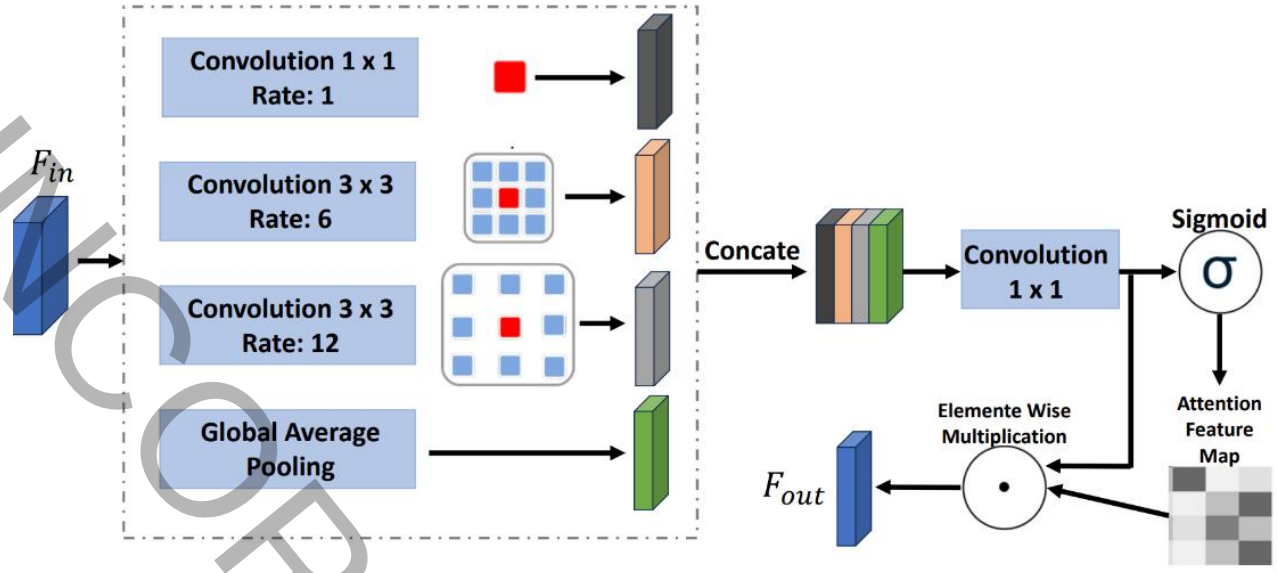


Fig. 6. Atrous Pyramid Attention Module (APAM) Components: Dilated Convolutions and Attention Mechanism.

aggregation, MBFF selectively emphasizes the most relevant information, improving segmentation precision. The refined features from MBFF are directly fed into corresponding decoders, enabling efficient reconstruction of segmentation masks at multiple scales. This ensures robust feature fusion and refinement, ultimately leading to improved segmentation performance. Figure 7 provides a visual breakdown of its structure and processing flow.

We adopt element-wise multiplication rather than addition or concatenation because it introduces a multiplicative, gating-like interaction between branches: a feature is emphasized only when it is jointly supported across the fused levels, which suppresses background responses that are inconsistent between scales and sharpens mask boundaries. We note that pure multiplicative fusion can, in principle, attenuate a feature that is strong in one branch but weak in another. In MBFF, this effect is mitigated by the 1×1 convolution, batch normalization, and ReLU applied immediately after fusion (Figure 7): the post-fusion projection learns to recombine the gated responses, allowing informative single-branch features to be recovered rather than discarded.

3- 6- Hybrid Loss

To ensure robust segmentation performance, we employ a hybrid loss function that integrates Dice loss and Focal loss, balancing both regional and pixel-wise supervision. This combination is particularly effective in handling class imbalance and enhancing the model's ability to capture fine details in masked face segmentation. Our model has four outputs, as shown in Figure 3. Thus, the whole segmentation loss can be described as:

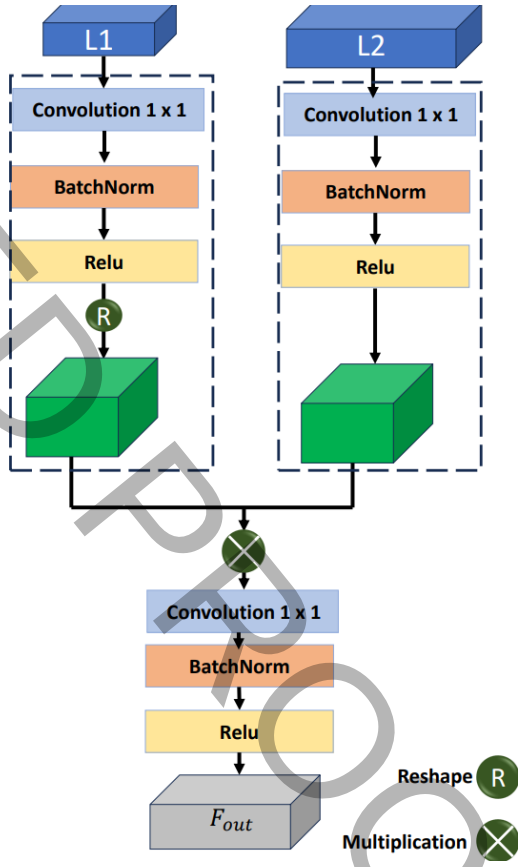


Fig. 7. The Multi-Branch Feature Fusion (MBFF) module integrates multi-scale features to enhance feature representation for precise masked face segmentation

UNCORRECTED PROOF

Algorithm 1. Atrous Pyramid Attention Module

Input: $X \in \mathbb{R}^{B\times C\times H\times W}$

Output: $Y_{\text{out}} \in \mathbb{R}^{B\times C\times H'\times W'}$

Step 1: Apply Dilated Convolutions

for $i = 1$ to n **do**

    $Y_i = f_{\text{dil}}(X; W_i, d_i)$

end for

Step 2: Compute Global Average Pooling

$Z = \frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W} X_{b,c,h,w}$

Z' = Broadcast(Z → H'×W')

Step 3: Concatenate Feature Maps

$Y = [Y_1, Y_2, \ldots, Y_n, \mathrm{Z}']$

Step 4: Apply 1x1 Convolution, BN, and ReLU

$Y_{\text{ASPP}} = \text{ReLU}(\text{BN}(f_{1\times1}(Y)))$

Step 5: Compute Attention Weights

$A = \sigma(f_{1\times1}(Y_{\text{APAM}}))$

Step 6: Apply Attention Weights

$Y_{\text{out}} = Y_{\text{APAM}} \odot A$

$$Loss = \sum_{n=1}^{N} Loss^{(n)} \quad (1)$$

$$Loss^{(n)} = \lambda \cdot Loss_{\text{Dice}}^{(n)} + (1-\lambda)\cdot Loss_{\text{Focal}}^{(n)} \quad (2)$$

where, $Loss^{(n)}$ is the side output loss for the n-th output, and N is the overall number of outputs, which is 4. The trade-off factor λ is employed to control the weight balance. By modifying the value range of λ, the relationship between similarity and category imbalance can be fine-tuned.

Dice Loss is a widely used metric in image segmentation tasks to measure the overlap between predicted and ground truth masks. It is derived from the Dice Coefficient, which is a similarity measure between two sets. The Dice loss is formulated as:

$$Dice\ \ Loss = 1 - \frac{2\sum_{i=1}^{N} G_i P_i}{\sum_{i=1}^{N} G_i + \sum_{i=1}^{N} P_i} \quad (3)$$

- G_i represents the ground truth value for the i -th pixel (manually labeled by an expert).
- P_i represents the predicted value for the i -th pixel (output of the segmentation model).
- N is the total number of pixels in the image.

The numerator $2\sum_{i=1}^{N} G_i P_i$ calculates the intersection (overlap) between the predicted mask and the ground truth.

The denominator $\sum_{i=1}^{N} G_i + \sum_{i=1}^{N} P_i$ represents the total number of pixels in both masks.

The final value ranges between 0 (perfect match) and 1 (no overlap).

We leverage the focal loss as a pixel-level loss to cope with low-confidence labels. It is defined as:

$$L_{binary-focal} = \begin{cases} -\alpha(1-y')^{\gamma}\log(y'), & y = 1 \\ -(1-\alpha)(y')^{\gamma}\log(1-y'), & y = 0 \end{cases} \quad (4)$$

where y is the ground truth and y' is the prediction. We set the α = 0.3 and γ = 2 based on this paper [10].

4- Experiments

In this section, we performed experiments to compare the performance of our proposed method with 9 state-of-the-art deep learning models for segmentation, which were selected as baseline approaches for evaluation of our proposed network.

UNCORRECTED PROOF

4- 1- Experimental Settings

During the training phase, we employ the Masked Face Segmentation Dataset (MFSD) for model evaluation [10]. All models are initialized using encoders pre-trained on ImageNet [35]. The intersection-over-union (IoU) metric is used to quantify the overlap between positive labels in a binary segmentation map, serving as a widely adopted benchmark for evaluating segmentation performance. All networks are implemented using the segmentation models library built on PyTorch [36]. Training is conducted with a batch size of 8. The optimization process is carried out using the Adam optimizer [37] with a learning rate set to 10^{-4}. The input images are resized to a resolution of 256 × 256 pixels before being fed into the model.

4- 2- Evaluation Metrics

To assess the performance of our network, we utilize multiple evaluation metrics, incorporating both conventional and task-specific measures for semantic segmentation. Below, we outline the five key metrics used for assessment:

Intersection over Union (IoU) is adopted as the principal metric for evaluating segmentation accuracy. In addition to IoU, we compute several other standard evaluation metrics [38].

$$Precision = \frac{TP}{TP+FP} \quad (5)$$

$$Recall = \frac{TP}{TP+FN} \quad (6)$$

$$F1 - score = \frac{2 \cdot TP}{2 \cdot TP+FP+FN} \quad (7)$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (8)$$

$$IoU = \frac{TP}{TP+FP+FN} \quad (9)$$

4- 3- Comparison with state-of-the-art models

On MSFD, we conduct a comparison of our network with other state-of-the-art approaches using the same training configuration of 500 testing images. Segmentation models are U-Net [39], Unet++ [40], Linknet [41], MAnet [42], PAN [43], DeepLabV3+ [44], BEDSN [45], fusionLungNet [46] and ELANet [47].

Quantitative Evaluation: As shown in Table 1, the evaluation metrics were computed and summarized for quantitative comparison. The best scores are highlighted in bold.

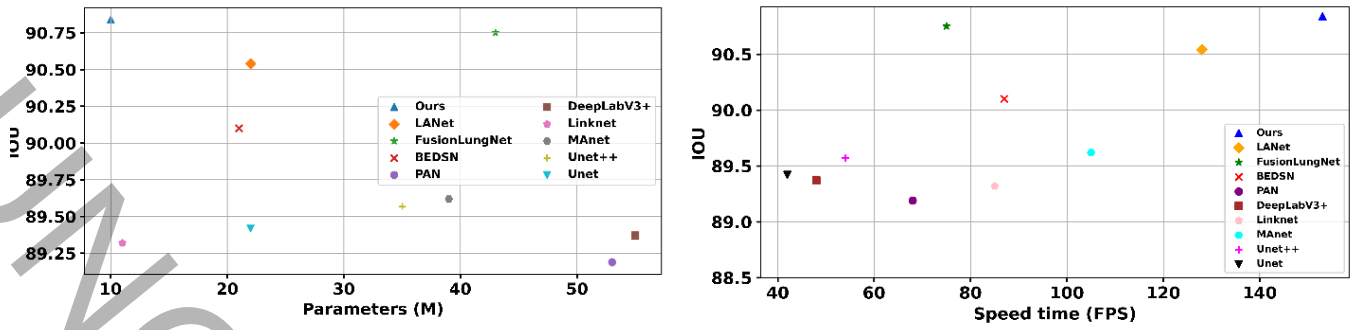
From the table, it can be observed that our proposed model achieved **90.84**, **96.47**, **94.81**, **98.2**, and **97.85** in terms of IoU, F1-score, Precision, Recall, and Accuracy, respectively. These results outperform other methods, demonstrating the effectiveness of our approach.

Compared to classical networks such as U-Net, U-Net++, and LinkNet, our model shows approximately 1.5%–2% improvement in IoU and F1-score, highlighting its superior segmentation capability. Furthermore, in comparison with

Table 1. A quantitative comparison of the results obtained on the MFSD test set using various image segmentation techniques.

Model	IoU	F1-score	Precision	Recall	Acc	Params(M)	FPS
Unet	89.42	94.21	91.60	96.98	96.24	22	42
Unet++	89.57	94.35	91.38	97.52	96.12	35	54
MAnet	89.62	94.44	91.49	97.58	97.25	39	105
Linknet	89.32	94.32	91.18	97.70	96.40	11	85
DeepLabV3+	89.37	94.43	91.64	97.40	96.99	55	48
PAN	89.19	94.18	91.35	97.20	96.03	53	68
BEDSN	90.10	95.63	93.52	97.84	97.52	21	87
FusionLungNet	90.75	96.27	**94.83**	97.77	**98.1**	43	75
ELANet	90.54	96.28	94.72	97.90	97.63	22	128
Ours	**90.84**	**96.47**	94.81	**98.2**	97.85	10	153

UNCORRECTED PROOF



(a) (b)

Fig. 8. Comparison of segmentation models in terms of parameter size and inference speed. (a) Parameter size (in million) versus IoU, showing the trade-off between model complexity and performance. (b) Inference speed (FPS) versus IoU, showing the efficiency of each model in real-time segmentation tasks.

more recent methods like BEDSN, FusionLungNet, and ELANet, our approach exhibits a marginal performance gain, reinforcing the suitability of the hybrid loss function for masked face segmentation.

To evaluate whether the improvement is statistically significant, we performed a paired t-test on the IoU scores obtained for all test images. The proposed method achieved a mean IoU of 90.63 ± 4.82, while FusionLungNet achieved 91.07 ± 4.55. The resulting p-value was 0.018 (<0.05), indicating that the improvement is statistically significant.

Additionally, our model achieves the highest FPS (**153**), indicating its efficiency in real-time applications, while maintaining a minimal parameter count (**10M**), making it significantly more lightweight compared to models like DeepLabV3+ and PAN. These findings suggest that our model effectively balances accuracy, computational cost, and inference speed. This trade-off between parameter size, inference speed, and segmentation performance is further visualized in Figure 8, where our model demonstrates an optimal balance of efficiency and accuracy compared to other methods.

Qualitative Evaluation: Figure 9 presents a qualitative comparison of various segmentation models for the task of masked face segmentation. The figure displays six random images from the dataset, their respective ground truth masks, and the segmentation output of our proposed model against ELANet, FusionLungNet, BEDSN, and U-Net++.

We can see from the figure that our proposed model produces segmentation masks that are very close to the ground truth, with clear boundaries and less noise.

Our model correctly segments the contours of the mask in the first and second images, whereas FusionLungNet and U-Net++ slightly over-segment the region of the mask along with a portion of the background. ELANet and BEDSN lose the smoothness of the mask boundary, leaving minor irregularities. In the third and fourth images, when the masks have complex textures, our model is more robust, retaining fine details that cannot be retained by other approaches. ELANet and BEDSN create small gaps in the estimated masks, whereas U-Net++ and FusionLungNet show small distortions along the edges of the masks.

In the fifth and sixth images, which feature occlusions and challenging lighting conditions, our model effectively separates the mask from the background, whereas FusionLungNet and BEDSN tend to produce noisier segmentations. ELANet and U-Net++ fail to maintain accurate mask contours, leading to minor segmentation errors. Overall, our proposed model consistently achieves more precise segmentations across various challenging scenarios, demonstrating its ability to handle different mask textures, occlusions, and background variations more effectively than competing methods.

4- 4- Ablation study

To further assess the effectiveness of MBFANet, we conducted ablation studies using our proposed dataset. These experiments specifically evaluate the impact of the CACR modules and APAM on segmentation performance.

We began with a baseline model that does not include any additional modules. Next, we integrated the CACR modules into this baseline to examine their contribution. Finally, we incorporated APAM into the network to observe its overall effect. The quantitative results, presented in Table 2, highlight the improvements at each stage, with the highest values bolded for clarity.

As shown in Table 2, our proposed MBFANet achieves the most significant improvements in IoU, F1-score, recall, and accuracy. This demonstrates that the CACR modules enhance the network's ability to detect subtle changes, while APAM effectively captures contextual information, leading to more accurate segmentation results.

UNCORRECTED PROOF

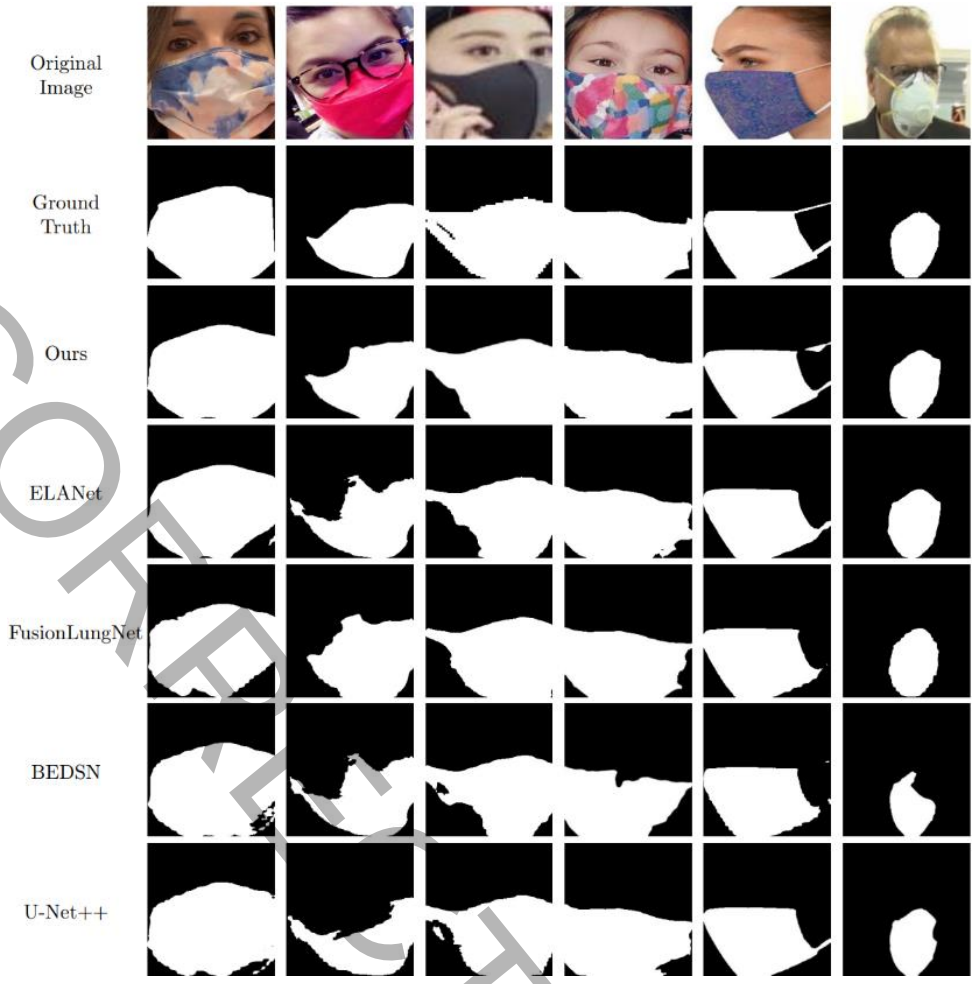


Fig. 9. Qualitative comparison of MFSD test set results with cutting-edge image segmentation algorithms. The first two rows display the original photos and the matching ground truth. Rows 3 to 8 illustrate the segmentation results derived from the proposed method, ELANet, BEDSN, FusionLungNet, and U-Net++. Red boxes indicate the fine distinction between the ground truth, BASNet, and the proposed method.

Table 2. The performance of different configurations of MBFANet.

Method	IoU	F1-score	Precision	Recall	Acc
baseline	89.35	96.3	95.33	97.28	96.25
baseline+CACR	90.5	**96.92**	**95.86**	98.00	97.73
baseline+APAM	89.71	96.34	95.37	97.34	96.62
baseline+ CACR +APAM	90.74	96.79	95.83	97.76	97.81
MBFANet	**90.84**	96.47	94.81	**98.20**	**97.85**

UNCORRECTED PROOF

Table 3. MBFANet performance with various input picture resolutions.

	Input size		
	128 × 128	256 × 256	512 × 512
IoU	90.47	**90.84**	90.61

Table 4. The performance of different optimizers.

Optimizer	IoU	F1-score	Acc
SGD	88.72	95.87	97.1
Adamax	89.1	96.43	97.51
RMSprop	89.48	96.35	97.73
Adam	**90.84**	**96.47**	**97.85**

Table 5. Effect of the hybrid-loss trade-off weight λ.

λ	0.3	0.5	0.7
IoU	89.53	**90.84**	90.34

To evaluate the stability of the network, we trained the MBFANet with 3 different input sizes, including 128 × 128, 256 × 256, and 512 × 512. Based on the data presented in Table 3, the input size of 256 × 256 obtains the best IoU.

To examine the sensitivity of the hybrid loss to its trade-off weight λ in Eq. (2), we trained MBFANet with three settings, $\lambda \in \{0.3, 0.5, 0.7\}$, while keeping all other hyperparameters fixed. As shown in Table 5, the balanced setting $\lambda = 0.5$, which assigns equal importance to the Dice and Focal terms, achieves the highest IoU of 90.84%. Reducing λ to 0.3 (Focal-dominant) lowers IoU to 89.53%, while increasing it to 0.7yields 90.34%. These results indicate that an overly Focal-weighted loss weakens region-level overlap, whereas a Dice-leaning loss recovers most of the performance but still falls short of the balanced configuration. Accordingly, we set $\lambda = 0.5$ in all experiments.

In this article, we use Adam as our optimizer. We also compare different major optimizers: Adam, Adamax, RMSprop, and SGD in Table 4. The Adam optimizer outperformed the other optimizers

5- Conclusion

In this work, we introduced MBFANet, a lightweight and efficient deep learning model for real-time masked face segmentation. The network integrates a multi-branch feature fusion strategy with attention mechanisms to enhance segmentation quality while maintaining low computational complexity. Built upon the RepVGG backbone, MBFANet incorporates two key modules: Channel Aggregation and Coordinate Refinement (CACR) to improve spatial and channel-wise feature representation, and the Atrous Pyramid Attention Module (APAM) to capture multi-scale contextual information. In the decoder, the Multi-Branch Feature Fusion (MBFF) module refines and integrates multi-scale features. To address class imbalance and improve boundary precision, we employed a hybrid loss combining Dice and Focal losses.

Extensive experiments on the Masked Face Segmentation Dataset (MFSD) show that MBFANet achieves the highest IoU, F1-score, and recall among the compared methods, together with the fastest inference speed (153 FPS) and a compact model size (10M parameters). Rather than surpassing

UNCORRECTED PROOF

all baselines on every individual metric, the strength of MBFANet lies in the balance it strikes across accuracy, inference speed, and parameter count: while a few methods are marginally higher on precision or accuracy and a few lightweight networks use fewer parameters, none matches MBFANet's overall trade-off between segmentation quality and real-time efficiency. Ablation studies further confirm the contribution of each proposed module to this balance.

Our study also has limitations. The evaluation was conducted on a single benchmark dataset, and the model has not yet been assessed on real-world footage captured under unconstrained conditions, such as variable lighting, motion blur, and diverse camera viewpoints. Its robustness in such settings therefore remains to be verified.

In future work, we plan to address these limitations along several concrete directions. First, we will evaluate MBFANet on real surveillance video with variable frame rates and uncontrolled illumination to assess its practical robustness. Second, we will develop a more compact variant through knowledge distillation, targeting fewer than five million parameters for deployment on resource-constrained edge devices. Finally, we intend to extend the framework toward occlusion-aware face recognition, exploring self-supervised pre-training to improve generalization across diverse occlusion types.

References

[1] A.M. Mansourian, N. Karimi Bavandpour, S. Kasaei, An Efficient Knowledge Distillation Architecture for Real-time Semantic Segmentation, AUT Journal of Modeling and Simulation, 55(1) (2023) 99–108.

[2] P. Molnar, AI in border control and migration: techno-racism and exclusion at digital borders, in Edward Elgar Publishing, 2024, pp. 307–322.

[3] M. Imani, Histogram-Based Shape and Textural Characteristics for Facial Emotion Recognition, AUT Journal of Modeling and Simulation, 51(1) (2019) 45–56.

[4] V.L. Raposo, Facial recognition AI technology in healthcare and the law, in Edward Elgar Publishing, 2024, pp. 41–56.

[5] Y. Zhuang, S. Li, M. Shifat-E-Rabbi, X. Yin, A.H.M. Rubaiyat, G.K. Rohde, Local sliced Wasserstein feature sets for illumination invariant face recognition, Pattern Recognition, (2025) 111381.

[6] m. Azimi Hemat, Fuzzy Content-Based Image Retrieval Speed-up Using the Multi-Agent Platform, AUT Journal of Modeling and Simulation, 54(1) (2022) 3–18.

[7] J. Liao, T. Guha, V. Sanchez, Self-supervised random mask attention GAN in tackling pose-invariant face recognition, Pattern Recognition, 159 (2025) 111112.

[8] Y. Chen, J. Li, S. Shan, M. Wang, R. Hong, From static to dynamic: Adapting landmark-aware image models for facial expression recognition in videos, IEEE Transactions on Affective Computing, (2024).

[9] K. Muhamada, U. Sudibyo, B. Widjajanto, A.A. Ojugo, et al., Exploring Machine Learning and Deep Learning Techniques for Occluded Face Recognition: A Comprehensive Survey and Comparative Analysis, Journal of Future Artificial Intelligence and Technologies, 1(2) (2024) 160–173.

[10] S. Rezvani, M. Fateh, H. Khosravi, ABANet: Attention boundary-aware network for image segmentation, Expert Systems, 41(9) (2024) e13625.

[11] E. Shabaninia, F. Shafizadegan, H. Nezamabadi-pour, A.R. Naghsh-Nilchi, Incorporating Transformer Networks and Joint Distance Images into Skeleton-driven Human Activity Recognition, AUT Journal of Modeling and Simulation, 56(1) (2024) 69–86.

[12] S. Rezvani, F. Soleymani Siahkar, Y. Rezvani, A. Alavi Gharahbagh, V. Abolghasemi, Single Image Denoising via a New Lightweight Learning-Based Model, IEEE Access, 12 (2024) 121077–121092.

[13] A. Fateh, Y. Rezvani, S. Moayedi, S. Rezvani, F. Fateh, M. Fateh, V. Abolghasemi, BRISC: Annotated dataset for brain tumor segmentation and classification, Scientific Data, (2026).

[14] W. Hariri, Efficient masked face recognition method during the COVID-19 pandemic, Signal, Image and Video Processing, 16(3) (2022) 605–612.

[15] S. Ge, C. Li, S. Zhao, D. Zeng, Occluded Face Recognition in the Wild by Identity-Diversity Inpainting, IEEE Transactions on Circuits and Systems for Video Technology, 30(10) (2020) 3387–3397.

[16] J. Dong, L. Zhang, H. Zhang, W. Liu, Occlusion-aware GAN for face de-occlusion in the wild, in, 2020, pp. 1–6.

[17] Y. Li, S. Liu, J. Yang, M.-H. Yang, Generative face completion, in, 2017, pp. 3911–3919.

[18] P.B.S. Varma, S. Paturu, S. Mishra, B.S. Rao, P.M. Kumar, N.V. Krishna, SLDCNet: Skin lesion detection and classification using full resolution convolutional network-based deep learning CNN with transfer learning, Expert Systems, 39(9) (2022) e12944.

[19] Q. Duan, L. Zhang, Look more into occlusion: Realistic face frontalization and recognition with BoostGAN, IEEE Transactions on Neural Networks and Learning Systems, 32(1) (2020) 214–228.

[20] N.U. Din, K. Javed, S. Bae, J. Yi, A novel GAN-based network for unmasking of masked face, IEEE Access, 8 (2020) 44276–44287.

[21] A. Fateh, M.R. Mohammadi, M.R.J. Motlagh, Msdnet: Multi-scale decoder for few-shot semantic segmentation via transformer-guided prototyping, arXiv preprint arXiv:2409.11316, (2024).

[22] H.-S. Yoon, S.-W. Park, J.-H. Yoo, Real-time hair segmentation using MobileUNet, Electronics, 10(2) (2021) 99.

[23] H.N. Tran, N.V. Nguyen, N.Q.P. Le, N.N.N. Nguyen, T.A.N. Le, V.D. Nguyen, Enhancing semantic scene segmentation for indoor autonomous systems using

UNCORRECTED PROOF

advanced attention-supported improved UNet, Signal, Image and Video Processing, 19(2) (2025) 190.

[24] J. Sun, J. Shen, X. Wang, Z. Mao, J. Ren, Bi-Unet: A Dual Stream Network for Real-Time Highway Surface Segmentation, IEEE Transactions on Intelligent Vehicles, 8(2) (2023) 1549–1563.

[25] R. Jiang, R. Chen, L. Zhang, X. Wang, Y. Xu, AM-MulFSNet: A fast semantic segmentation network combining attention mechanism and multi-branch, IET Image Processing, 18(7) (2024) 1733–1744.

[26] X. Song, X. Fang, X. Meng, X. Fang, M. Lv, Y. Zhuo, Real-time semantic segmentation network with an enhanced backbone based on Atrous spatial pyramid pooling module, Engineering Applications of Artificial Intelligence, 133 (2024) 107988.

[27] W. Wang, B. Pan, Y. Ai, G. Li, Y. Fu, Y. Liu, LightCM-PNet: A lightweight pyramid network for real-time prostate segmentation in transrectal ultrasound, Pattern Recognition, 156 (2024) 110776.

[28] Q. Zhao, H. Fang, Y. Pang, G. Zhu, Z. Qian, PL-UNet: a real-time power line segmentation model for aerial images based on adaptive fusion and cross-stage multi-scale analysis, Journal of Real-Time Image Processing, 22(1) (2025) 39.

[29] X. Hu, J. Feng, J. Gong, LFFNet: lightweight feature-enhanced fusion network for real-time semantic segmentation of road scenes, Pattern Analysis and Applications, 27(1) (2024) 27.

[30] L. Chen, L. Gu, D. Zheng, Y. Fu, Frequency-adaptive dilated convolution for semantic segmentation, in 2024, pp. 3414–3425.

[31] J. Huang, H. Ren, S. Liu, Y. Liu, C. Lv, J. Lu, C. Xie, H. Lu, Real-time Attentive Dilated U-Net for Extremely Dark Image Enhancement, ACM Transactions on Multimedia Computing, Communications and Applications, 20(8) (2024) 1–19.

[32] Q. Liu, C. Wang, Z. Li, Y. Qi, J. Fang, Attention-based lightweight asymmetric network for real-time semantic segmentation, Engineering Applications of Artificial Intelligence, 130 (2024) 107736.

[33] G. Zhang, Y. Lu, X. Jiang, F. Yan, M. Xu, Context-Aware Adaptive Weighted Attention Network for Real-Time Surface Defect Segmentation, IEEE Transactions on Instrumentation and Measurement, 73 (2024) 1–13.

[34] X. Ding, X. Zhang, N. Ma, J. Han, G. Ding, J. Sun, RepVGG: Making VGG-style convnets great again, in 2021, pp. 13733–13742.

[35] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., Imagenet large scale visual recognition challenge, International journal of computer vision, 115(3) (2015) 211–252.

[36] P. Yakubovskiy, Segmentation Models PyTorch, in GitHub, 2020, pp. Available from: https://github.com/qubvel/segmentation_models.pytorch.

[37] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980, (2014).

[38] P.O. Editors, Retraction: Skin cancer segmentation and classification by implementing a hybrid FrCN-(U-Net) technique with machine learning, in Public Library of Science, San Francisco, CA, USA, 2026.

[39] O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional networks for biomedical image segmentation, in, 2015, pp. 234–241.

[40] Z. Zhou, M.M. Rahman Siddiquee, N. Tajbakhsh, J. Liang, Unet++: A nested U-Net architecture for medical image segmentation, in Springer, 2018, pp. 3–11.

[41] A. Chaurasia, E. Culurciello, LinkNet: Exploiting encoder representations for efficient semantic segmentation, in, 2017, pp. 1–4.

[42] T. Fan, G. Wang, Y. Li, H. Wang, Ma-net: A multi-scale attention network for liver and tumor segmentation, IEEE Access, 8 (2020) 179656–179665.

[43] H. Li, P. Xiong, J. An, L. Wang, Pyramid attention network for semantic segmentation, arXiv preprint arXiv:1805.10180, (2018).

[44] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, in 2018, pp. 801–818.

[45] X. Li, L. Xie, C. Wang, J. Miao, H. Shen, L. Zhang, Boundary-enhanced dual-stream network for semantic segmentation of high-resolution remote sensing images, GIScience \& Remote Sensing, 61(1) (2024) 2356355.

[46] S. Rezvani, M. Fateh, Y. Jalali, A. Fateh, Fusionlungnet: Multi-scale fusion convolution with refinement network for lung CT image segmentation, arXiv preprint arXiv:2410.15812, (2024).

[47] Q. Yi, G. Dai, M. Shi, Z. Huang, A. Luo, ELANet: effective lightweight attention-guided network for real-time semantic segmentation, Neural Processing Letters, 55(5) (2023) 6425–6442.

HOW TO CITE THIS ARTICLE

Z. Rawan Abdulraheem, M. Fateh, S. Ferdowsi, S. Rezvani, MBFANet: A Lightweight Network with Multi-Branch Fusion and Attention Modules for Real-time Segmentation, AUT J. Model. Simul., 58(2) (2026) 163-178.

DOI: 10.22060/miscj.2026.25626.5483

UNCORRECTED PROOF

UNCORRECTED PROOF