A Hybrid EfficientNet and Self-Attention Architecture for Masked Face Recognition

Lekha Prajapati[1], Girish Katkar[2], Ajay Ramteke[3]

[1]Research Scholar, Department of Computer Science, Taywade College Koradi, (M.S.), India.

[2,3]Assistant Professor, Department of Computer Science, Taywade College Koradi, (M.S.), India.

DOI: https://dx.doi.org/10.51584/IJRIAS.2026.110200035

Received: 10 February 2026; Accepted: 16 February 2026; Published: 2 March 2026

ABSTRACT

The use of facial masks in the real-world setting has made the Masked Face Recognition (MFR) a critical research problem in Pattern Recognition. The classical face recognition technology is highly impaired in performance when the nose and mouth are covered as the main facial features. In this paper, a strong hybrid deep learning model will be suggested, which integrates the EfficientNet-B0 convolutional neural network with a selfattention mechanism to promote the learning of discriminative features on a partially visible part of the face. EfficientNet-B0 is also an efficient and scalable feature extractor, and the self-attention module allows global contextual reasoning and adaptable attention to unoccluded areas of the face, especially the periocular area. The suggested model is tested on the actual MFR2 dataset and has a recognition rate of 0.99, which proves to be better than the traditional CNN-based methods. The experimental test proves that the combination of selfattention can greatly enhance the resilience to the obstruction of the object by obstructing the features of the mask. The findings suggest that the hybrid architecture proposed is quite appropriate in real-time biometric authentication and surveillance and access control systems with masked environments.

Keywords: Masked face recognition, Self-attention mechanism, EfficientNet-B0, Hybrid deep learning, MFR2 dataset.

INTRODUCTION

The COVID-19 pandemic has radically changed the quality of the biometric authentication systems and posed more challenges to face recognition technologies than ever. The universal use of facial mask as a protective health tool has generated a pressing demand of high-quality masked face recognition systems that can retain high accuracy even in the cases of considerable face masking. The classic face recognition systems that are based on the features obtained in the lower part of the face such as nose, mouth and chin, are severely affected by the masking of these areas, which results in severe performance reduction [1], [2], [3]. This corruption is a major threat to the security systems, access control systems and identity verification programs which have become part of the contemporary infrastructure.

Recent developments in deep learning have triggered a massive development in solving the masked face recognition problem. Convolutional Neural Networks (CNNs) have been shown to be extremely successful at extraction of hierarchical visual features of face images [4], [5]. Nevertheless, CNNs are also characterized by intrinsic weaknesses in capturing long-range dependencies and global contextual information, which are especially important in the cases when large parts of the face are covered [6], [7]. The recent advent of transformer architectures, which were initially created to support natural language processing, has facilitated new possibilities of computer vision tasks by allowing the representation of global dependencies effectively with the help of self-attention mechanisms [8], [9], [10].

In spite of these, current methods of masked face recognition have a number of underlying constraints. Pure CNN-based techniques have problems with adapting to visible areas of the face and frequently cannot achieve the higher-order spatial correlation required to make fine distinctions between faces in cases when significant facial features are hidden by the hand [11], [12]. Transformer-based models are good at capturing global context, but can be inductively biased to effectively extract features when using high-resolution facial images, and can overfit when training data is small [13], [14]. Moreover, the majority of known techniques fail to effectively deal with the regularization issues of learning with partially blocked data, in which the threat of overfitting to spurious correlations is highly increased. The latest developments in the field of deep learning have facilitated more powerful feature extraction using Convolutional Neural Networks (CNNs). But CNNs are local by nature and can potentially be trained to pick up features in the mask itself, resulting in poor generalization[15]. To overcome this weakness, attention mechanisms have become of great importance in that they enable the models to dynamically direct attention to the most informative part of an image.

In this study, a hybrid network that integrates EfficientNet-B0 and a self-attention mechanism is suggested to masked face recognition. The combination of efficientNet-B0 and the self-attention module offers a good balance between computational efficiency and representational power, and global contextual understanding. The suggested model is trained and tested on the MFR2 dataset, and the accuracy of the hybrid approach is 99, which proves the effectiveness of the hybrid approach.

Related Work

This part is a review of the vast amount of literature on masked face recognition, transformer-based face recognition and hybrid architectures. We are mentioning the development of strategies and defining the gaps that are occupied by our work.

CNN-Based Masked Face Recognition

In the last ten years, Convolutional Neural Networks have served as the paradigm of face recognition and have reached impressive performance on unconstrained face recognition benchmarks. This has been however shown to be the fundamental limitation of pure CNN approaches, with the advent of masked face recognition as a critical issue.

New CNN-based approaches have involved the use of more advanced attention schemes. Wan et al. [16] suggested a two-branched convolutional self-attention network, which is a hybrid of CNN feature extractor and self-attention modules.

They combine multi-head self-attention with a convolutional framework in their modified convolutional self-attention module (MCSAM), which yields better performance on masked face recognition. Nevertheless, this method remains largely based on convolutional operations and is not utilizing transformer architectures to the fullest. Wan et al.

[17] followed this direction in a follow-up work, but with knowledge distillation, where an instructor-learner model is used to enhance the performance of a convolutional self-attention network.

Ge et al. [18] presented a Convolutional Visual Self-Attention Network (CVSAN) which adds self-attention blocks to the convolutional features. Although such a hybrid method has potential, it uses self-attention as a supplementary mechanism, not as a fundamental architecture, and thus it cannot effectively model global dependencies.

The initial research on occluded face recognition concentrated on strong feature extraction and loss functions that are aware of occlusions. The approach Qiu et al. [19] suggested includes end-to-end training, masking the corrupted features, and enhances the resistance to occlusions.

Their algorithm learns how to detect and remove features of occluded regions, but uses CNN architectures that are poor at learning long-range interactions between visible facial parts. Wang et al.

[20] presented DSA-Face that uses multiple and sparse attentions to identify discriminative parts of the face and subdues distracted areas. Although efficient, this method relies on attention mechanisms in a CNN structure and does not have the global modeling ability of transformers.

Transformer-Based Face Recognition

General computer vision use of Vision Transformers (ViT) has encouraged their use in face recognition. Transformers have special benefits to masked face recognition by being able to learn long-range dependencies, and attend to informative regions.

Zhao et al. [21] introduced the Masked Face Transformer (MFT), which is based on Swin Transformer and proposes a Masked Face-compatible Attention (MFA) mechanism. MFA increases the range of attention by adding more window partition settings and inhibits communication between masked and unmasked areas. Another ClassFormer module is also presented in the paper to increase the intra-class aggregation. MFT shows both state-of-the-art results on simulated and real masked face datasets, confirming that transformer architectures are effective in this task.

Zhu et al. [22] created FaceT, which is a holistic and masked face recognition framework built on Vision Transformers and prompt-based strategies.

They tackled the problem of training ViTs in isolation by using an initializing model parameters with a patch reconstruction proxy task, which found better convergence and performance.

FaceT matches or is more accurate than state-of-the-art CNNs on both holistic and masked face recognition tasks, which illustrates the generality of transformer-based architectures.

Wang et al. [23] discovered learning 3D face representations using Vision Transformers to mask face recognition. Their method breaks whole and half faces into sequence images and uses the transformer capability to extract the relationship between the slices of the image to fill the lost face details.

Using geometric properties of 3D face point clouds, they made significant improvements over 2D methods and accuracy increased by 9.86% on Bosphuras, 16.77% on CASIA-3D Facev1, 2.32% on StirlingESRC and 34.81% on Ajmal main database.

Ouannes et al. [24] examined the use of Vision Transformers in face recognition under degraded conditions such as partial occlusions, variations in lighting, and pose variations. They use a transformer encoder to obtain discriminative features and transformer decoder with self-attention mechanisms to refine representations, as shown to be optimal even in degradation conditions.

A full-scale comparison of Vision Transformers and CNNs in face recognition tasks was also performed by Rodrigo et al. [25], where the authors evaluated their results on five datasets such as the Labeled Faces in the Wild, Real World Occluded Faces, and Surveillance Cameras Face.

Their results show that Vision Transformers both achieve better accuracy and resilience to distance and occlusions compared to CNNs and have smaller memory footprints and are also faster to run.

Attention Mechanisms for Occluded Face Recognition

The attention mechanisms have become a very important element in the handling of occlusions in face recognition in which networks can focus on the visible and informative parts of the face in an adaptive manner.

Wang et al. [26] proposed AAN-Face that utilizes attention augmented networks to recognize faces. Although they were not explicitly constructed to deal with masked face, their attention systems offer some information on how networks can be trained to focus on discriminative facial features.

Tiong et al. [27] came up with a Flexible Biometric Recognition (FBR) system which employs Multimodal Fusion Attention (MFA) in fusing face and periocular biometrics[28]. Their Multimodal Prompt Tuning (MPT) system facilitates cross-modality interactions, yet retains the specific qualities, and has shown outstanding results on intra- and cross-modality recognition tasks on four standard datasets.

Phan et al. [29] suggested a new 2-image Vision Transformer model that compares the images on the patch level with cross-attention. Their model trained on 2M pairs of CASIA Webface achieves a similar accuracy to DeepFace-EMD on out-of-distribution data with the added benefit of being more than two times faster, and their evaluation protocols and datasets have been used to compare methods. Their contribution emphasizes the importance of regular methodologies of evaluation in this fast changing sphere.

EfficientNet-B0 ARCHITECTURE

EfficientNet is a series of convolutional neural networks that maximize performance at a minimum cost of computation.

Compared to traditional architectures where depth, width, or resolution is scaled independently, EfficientNet uses compound scaling where all three dimensions are scaled equally using constant coefficients. EfficientNet-B0 is the base model in the EfficientNet family.

It also uses Mobile Inverted Bottleneck Convolution (MBConv) blocks, which consist of depthwise separable convolutions and squeeze-and-excitation modules. These components allow efficient feature extraction with fewer parameters and are able to maintain representational capacity.

The mathematical definition of this compound scaling method is that a principal coefficient, φ, uniformly scales the depth (d), width (w) and resolution (r) of the network using the equations below:

$$d = \alpha\text{^}\varphi,\ w = \beta\text{^}\varphi,\ r = \gamma\text{^}\varphi \qquad (1)$$

subject to: $\alpha \cdot \beta^2 \cdot \gamma^2 \approx 2,\ \alpha, \beta, \gamma \geq 1$

Depth (d) Refers to the depth of the network. Adding depth enables the model to capture more complicated attributes but may cause vanishing gradients, which is overcome by the EfficientNet structure.Width (w) Refers to the count of channels (filters) in each layer.

A broader network is able to capture finer-grained features, e.g. the texture of a mask, the exact shape of a periocular area.Resolution (r) Refers to the size of the input image. Resolution is also increased to enable the model to view smaller and more detailed patterns in the visible parts of the face.

EfficientNet-B0 is effective in the context of masked face recognition to obtain fine-grained texture information in exposed areas of the face.

The squeeze-and-excitation operation also benefits the channel-wise feature importance, which enables the network to emphasize the informative features in a more adaptive manner prior to the higher-level processing, when it is fine-tuned on the MFR2 dataset.

Self-Attention Core Mechanism of Transformer

Transformer architecture is a type of neural network that was created to work with sequential data by applying attention mechanisms rather than recurrence or convolution. It finds extensive application in NLP, and in more vision tasks. Multi-Head Self-Attention (MHSA) The model can compute Query (Q), Key (K), and Value (V) matrices to capture global dependencies in the input sequence.Figure 1 represents the architecture of the Transformer, where the main idea of modeling global dependencies in the input sequence is self-attention.

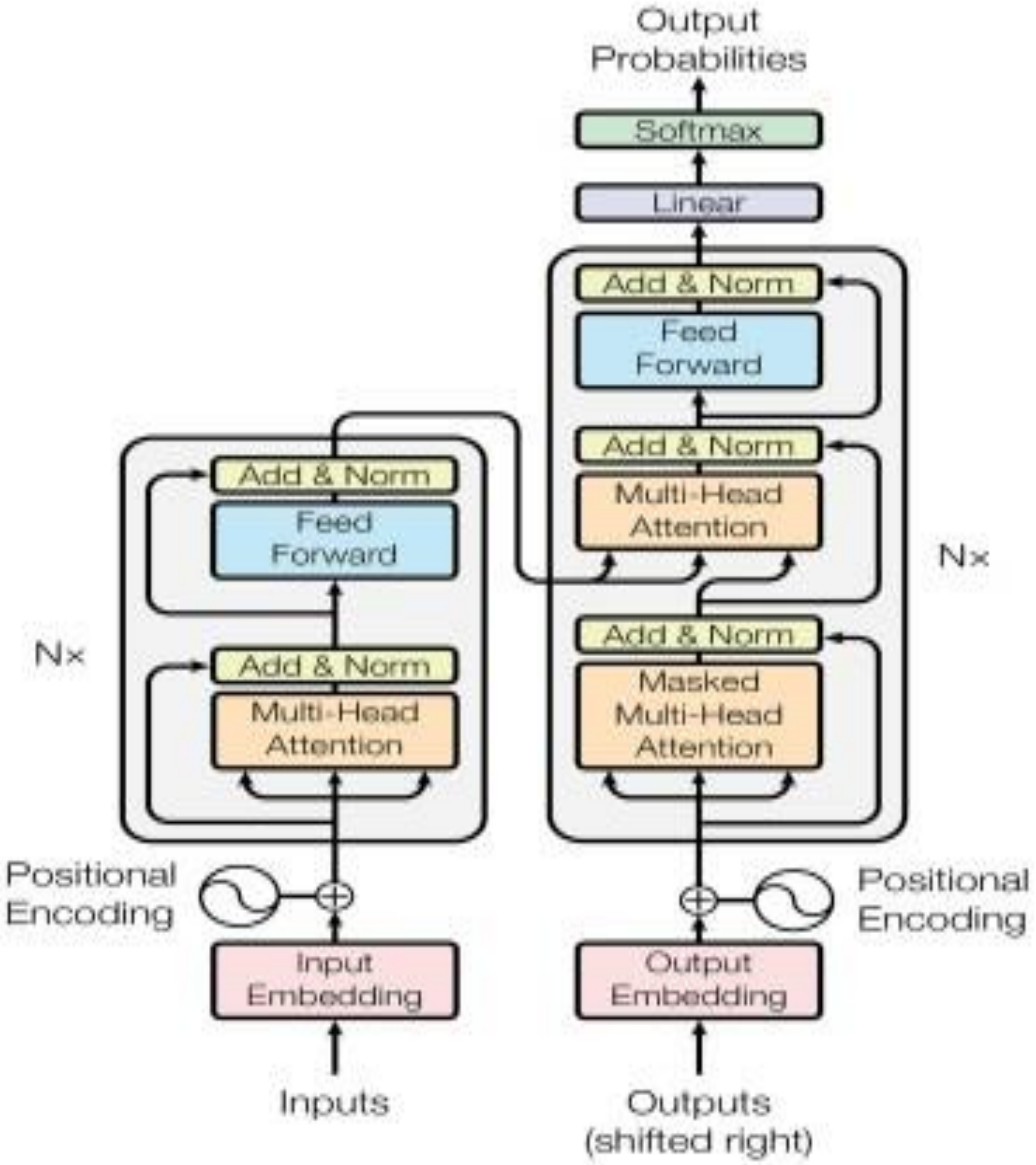
1. Multi-head self-attention is used in the encoder to enable a rich contextual feature learning by making each input element attend to all other elements.

2. The decoder makes use of masked self-attention such that predictions are based solely on the output that has been generated before.

3. Normalization of layers and residual connectivity stabilizes the training and promotes information flow within the layers.

4. The last two layers are the linear and softmax layers which transform the attention-enhanced representations into output probability distribution.

Figure 1. Transformer Architecture



Although CNNs are effective in local pattern learning, they are weak in global relationship learning between spatial locations. This is a weakness especially in masked face recognition where the model has to learn to disregard covered areas and concentrate on the visible features.

The first stage of the self-attention procedure implies projecting the input feature map X, which is gained by the EfficientNet-B0 backbone, into three separate latent spaces. Such linear transformation is expressed as:

$$Q = XW^{Q}, K = XW^{K}, V = XW^{V} \qquad (2)$$

e Input (X) Here, X denotes the spatial feature map of the EfficientNet-B0 backbone. The feature map is rearranged into a sequence of tokens, each token being a spatial position on the face.The Learnable Weights

(W^{q}, W^{k}, W^{v}) These are learnable weight matrices that are learned during training with the Adam optimizer. They allow the network to identify the most relevant feature to use in determining the identity.The Query (Q) Represents the existing feature token that requests relevant information in other spatial features within the picture.The Key (K) Represents a set of reference tags of all the spatial features that the Query can evaluate the importance of each feature location.The Value (V) Holds the actual visual information of the facial features. When a high relevance is established between a Query and a Key the identity-preserving information corresponding to this matrix is obtained by converting the raw convolutional features into Q, K, and V representations.The conversion of the raw convolutional features into Q, K, and V representations transforms the local feature analysis of the facial structure into a global interpretation. This is a global reasoning ability that is of particular importance in masked face recognition, whereby the lower part of the face is either partially or wholly covered.

After the Query, Key and Value representations have been generated, the core attention mechanism makes a weighted sum of the values. The relevance of various facial regions is determined in this process by computing the compatibility of queries and keys defined by the Scaled Dot-Product Attention mechanism:

$$\text{Attention}(Q, K, V) = \text{softmax}(\ (Q \cdot K^T) / \sqrt{d_k}\ ) \cdot V \qquad (3)$$

The Dot Product ($Q\ K^T$) This is used to determine the similarity of each query and every key. In masked face recognition it enables the model to estimate the strength of certain facial feature, e.g. a region about the eyes, in relation to other identity preserving areas of the face. The Scaling Factor ($\sqrt{d_k}$) Here, d_k is the dimensionality of the key vectors. Dividing by the square root of d_k division is also a significant normalization technique that avoids the values of the dot product of becoming too large. This stabilization prevents very small gradients in the softmax function and guarantees stable and effective training.The softmax function transforms the similarity scores into a normalized probability distribution (so-called attention weights), which add up to one. The process is effective to suppress noise presented by the masked regions of the face by assigning them low weights whereas identity-relevant information is promoted by the model by multiplying the attention weights with the Value matrix (V). This discriminative ability improves the discriminative power of the network.This attention mechanism allows the hybrid framework to be robust to various mask types and occlusion patterns since the attention weights are dynamically adjusted to a set of facial regions that are in view.

In order to increase the representational capacity of the network further, the Multi-Head Self-Attention (MHSA) is applied in the proposed architecture. This is a mechanism, which enables the model to concurrently process the information of various subspaces of the representation at various spatial locations. MHSA operation can be said to be as follows:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots , \text{head}_h)\ W^O \qquad (4)$$

where: $\text{head}_i = \text{Attention}(Q\ W_i$٩, $K\ W_i^k, V\ W_i^v)$

Parallel Processing (head 3) As an alternative to a single attention operation, the Query, Key and Value matrices are projected into h parallel attention heads. One attention head can be used to attend to multiple identitydiscriminative regions of the face (e.g. eyebrow contours, eye shape, forehead texture) simultaneously.Subspace Learning Each head of attention is represented by independent learnable projection matrices (W_i٩, W_i^k, W_i^v).

This allows the model to acquire a variety of feature associations in a variety of representation subspaces. As an example, one of the heads of attention can be more focused on spatial geometry, another can focus on finegrained texture detail.The output of all attention heads is then concatenated along the feature dimension, effectively combining the multiple views of the facial representation into a single unified feature representation.Linear Projection (W W) A final learnable projection matrix W W is then applied to the concatenated output to revert the feature dimensionality needed by the other layers, so as to ensure that the refined representation is appropriate in the final classification stage. Although some facial features might be obscured by mask straps or changes in lighting, other heads will be able to retrieve enough discriminative information using the unoccluded parts of the periocular.

A residual connection with layer normalization is used to ensure that the training process remains stable, and the features are not degraded as they flow through the attention layers.

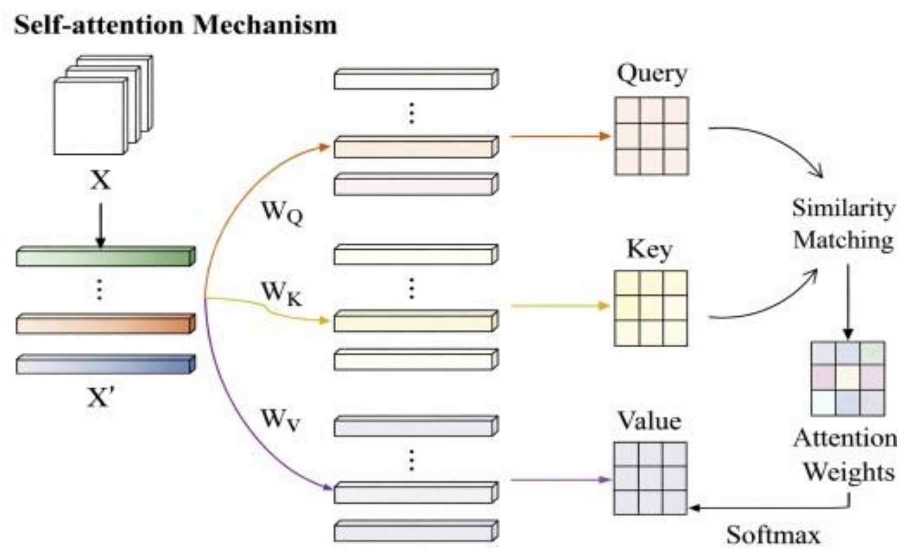
$$Z = \text{LayerNorm}(\ X + \text{SelfAttention}(X)\ ) \qquad (6)$$

Residual Connection (X + SelfAttention(X)) This is the operation, also known as a skip connection, which includes the original input feature X to the output of the self-attention module. This architecture enables a more gradient-flowing backpropagation, where the identity-discriminative features extracted by the EfficientNet-B0 backbone are not lost and distorted in an excessive manner by the self-attention mechanism.Feature Integration The residual addition is to make the final representation Z integrate high-quality local texture features learned by the convolutional layers and global contextual relationships learned by the self-attention mechanism. The combination of these two improves the strength in the face mask recognition.Layer Normalization re-centers and re-scales the feature representations on the hidden dimensions. ConvergenceGradient StabilityCombined, the residual connection and normalization of layers effectively reduce internal covariate shift which causes the vanishing gradient problem and makes training stable and efficient. This stabilization is important in ensuring that the model reaches high recognition accuracy of 99% without training instability.

The proposed hybrid framework is able to utilize this stabilization layer to fine-tune the pretrained EfficientNet weights in the masked face recognition domain and achieve high generalization performance.

The self-attention mechanism can solve this problem by calculating the relationships between all the positions in the spatial positions of the feature map. The extracted feature map of EfficientNet-B0 is rearranged into a sequence representation, and each spatial location is considered a token. The interactions between these tokens are then modeled by using multi-head self-attention, which learns to emphasize more the parts of the body (eyes and forehead) and suppress the parts of the body (mask). In stable training and gradient propagation, residual connections and layer normalization are used. This refinement, based on attention, substantially increases the discriminative power of the obtained features, and makes the model resilient to changes in the mask type and the severity of occlusions.

Figure 2. Self -Attention Mechanism



The self-attention mechanism in Figure -2 In which the input feature map is converted to query, key, and value representations with learnable weight matrices.The similarity between the queries and keys is calculated to compute the relevance of various spatial features.This similarity score is normalized using a softmax function to generate attention weights.The attention weights are applied to the value vectors to emphasize important features at the expense of less important features.This attention-based refinement is particularly effective in its ability to focus on identity-preserving features of the face, which is especially effective at masked face recognition.

Proposed Hybrid Experimental Approach

The proposed masked face recognition model employs EfficientNetV2B0 as the backbone feature extractor, initialized using pre-trained ImageNet weights. The top classification layer of the original architecture is excluded to allow integration of a task-specific classification head. By setting base_model.trainable = True, all layers of the pre-trained backbone are fine-tuned during training. This enables domain adaptation, allowing the network to learn discriminative representations tailored to masked facial characteristics rather than relying solely on generic ImageNet features.

The output feature map of the backbone network be represented as:

$$F \in \mathbb{R}^{B\times H\times W\times C} \qquad (7)$$

where Bdenotes the batch size, $H$and Wrepresent spatial dimensions, and Ccorresponds to the number of feature channels.

To incorporate the self-attention mechanism, the 4D feature map is reshaped into a 3D tensor:

$$F' \in \mathbb{R}^{B\times(H\cdot W)\times C} \qquad (8)$$

In this representation, each spatial location $(H \cdot W)$is treated as an independent token, enabling the attention mechanism to model long-range dependencies across facial regions. This transformation allows the network to capture contextual relationships between visible facial features, which is particularly important in masked face recognition scenarios where significant portions of the lower face are occluded.

A custom attention block is then applied to the reshaped tensor. The self-attention mechanism computes query (Q), key (K), and value (V) projections and refines feature representations using:

$$\mathrm{Attention}(Q, K, V) = \mathrm{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (9)$$

where $d_k$is the dimensionality of the key vectors. This operation enhances global feature interaction and improves discriminative capability under occlusion conditions.

Following attention refinement, the tensor is reshaped back to its original 4D structure:

$$\mathbb{R}^{B\times H\times W\times C} \qquad (10)$$

A GlobalAveragePooling2D layer is subsequently applied to reduce spatial dimensions and generate a compact feature vector:

$$z_c = \frac{1}{H\cdot W}\sum_{i=1}^{H}\sum_{j=1}^{W} F_{i,j,c} \qquad (11)$$

This operation aggregates spatial information while preserving channel-wise discriminative features.

To reduce overfitting, a Dropout layer with a rate of 0.4 is introduced after global pooling. During training, 40% of neurons are randomly deactivated, which prevents co-adaptation among feature detectors and enhances generalization performance on unseen masked face samples.

The final classification stage consists of a fully connected Dense layer with Softmax activation, where each output neuron corresponds to a distinct identity class.

The Softmax function converts logits into normalized class probabilities:

$$P(\, y = i \mid x \,) = \frac{e^{z_i}}{\sum_{j=1}^{N} e^{z_j}} \qquad (12)$$

where Ndenotes the number of identity classes.

The complete model is constructed by connecting the EfficientNet input tensor to the custom classification head. The network is compiled using the Adam optimizer with a learning rate of 1×10^{-4}, ensuring adaptive gradientbased optimization. The categorical cross-entropy loss function is employed for multi-class identity classification, and accuracy is used as the primary evaluation metric.

Optimization Strategy

To ensure robust convergence and prevent overfitting, three standard Keras callbacks are incorporated into the training pipeline. A ModelCheckpoint callback monitors validation accuracy at the end of each epoch and persistently saves the model weights only when an improvement is observed. This guarantees that the final saved model corresponds to the epoch of peak generalization performance, rather than the last epoch of training. An Early Stopping callback monitors validation loss with a patience window of five epochs. If no improvement in validation loss is detected within this window, training is terminated prematurely to prevent overfitting and unnecessary computation. Upon stopping, the model weights are automatically restored to those corresponding to the best-performing epoch. A ReduceLROnPlateau callback dynamically adjusts the learning rate during training. Specifically, if validation loss fails to decrease over three consecutive epochs, the learning rate is scaled by a reduction factor of 0.3. This adaptive scheduling allows the optimizer to escape shallow local minima and achieve finer convergence as training progresses.

Augmentation Strategies

To improve the generalization capability of the model and mitigate the risk of overfitting, a comprehensive data augmentation strategy was applied exclusively to the training set during preprocessing. All pixel values were first normalized to the range [0, 1] by rescaling with a factor of 1/255, ensuring numerical stability during gradient-based optimization. Random horizontal flipping was incorporated to introduce mirror-invariant representations, while random rotations of up to 15 degrees were applied to account for moderate orientational variability in the input images.

Additionally, random zoom augmentation of up to 20% was employed to simulate scale variations, and random horizontal and vertical shifts of up to 10% of the total image width and height, respectively, were applied to encourage spatial translation invariance. Together, these transformations artificially expand the effective diversity of the training distribution, reducing the model's tendency to memorize specific training examples. In contrast, no augmentation was applied to the test set beyond the same pixel normalization, ensuring that evaluation was performed on unmodified images representative of real-world input conditions.

Identity Recognition Under Mask Occlusion

In contrast to traditional face recognition models which are highly dependent on the entire geometry of the face, the suggested algorithm is centered on identity-conserving parts of the face. EfficientNet-B0 identifies hierarchical features of all visible regions of the face such as eyes, eyebrows, and forehead. These areas are not much blocked in masked situations and they have enough discriminative information to recognize identity.

Role of Self-Attention in Identity Discrimination

Attention mechanism is important in the identification of masked face identity as it allows the model to prioritize the uncovered areas of the face, Suppress features that are related to the mask area and capture the long-range feature between spatial facial features.Dynamically re-weighting the spatial features, the attention mechanism guarantees that information that is relevant to identity is dominant in the final representation even in case large parts of the face are covered with a mask.

Class-Level and Identity-Level Learning

The identity labels are used to train the algorithm in a supervised manner. One output class is associated with a single person in the dataset. Consequently, the softmax classifier will learn to directly regress attention-refined feature representations to individual identities, but not just to masked or unmasked categories. This combined class-level and identity-level learning approach allows the model to identify the same individual in both masked and unmasked conditions, thus enhancing generalization and accuracy of recognition of a masked face recognition task in the real world.

RESULTS AND DISCUSSION

Fig 3 show the training and validation accuracy of the proposed hybrid EfficientNet- Self Attention model over the various training epochs of the MFR2 dataset.Both curves keep improving progressively and level off at 0.98-0.99. The high correlation between training and validation accuracy during the training process shows that it has a good generalization ability and there is very little overfitting. Minor variations in subsequent eras are anticipated in the process of fine-tuning and do not point to deterioration of performance. The validation accuracy is always high, which reaffirms how effective the hybrid architecture is in the recognition of masked and unmasked faces.

Figure 3. Training and Validation Performance Graph

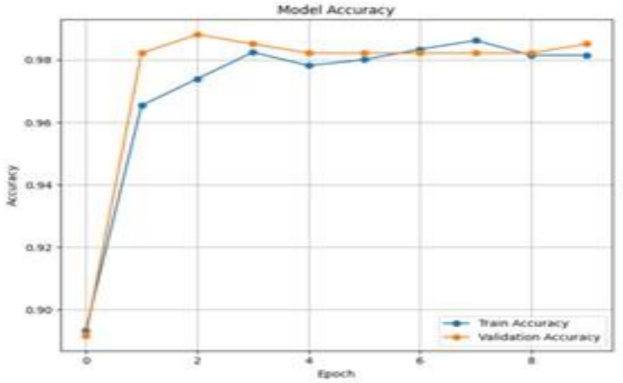


Figure 4. Loss Curves Graph

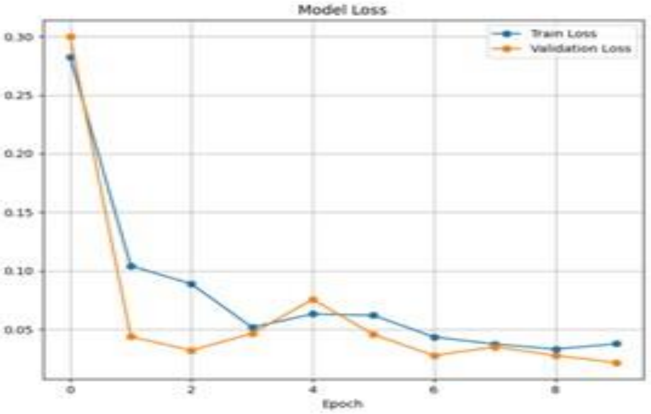


Figure 4 also confirms the effectiveness of the proposed model by the loss curves. First, training and validation losses are quite high at about 0.28 0.30, which is expected in the initial stages of training. The loss significantly reduces after the first epoch, indicating the rapid convergence and successful optimization, and as the training progresses the loss decreases steadily and reaches the values under 0.05 whereas the validation loss shows the same downward trend and attains the values near 0.02. The training

and validation loss curves are close to each other indicating a stable learning behavior and proving that the model is not overfitted.

The joint analysis of accuracy and loss curves shows that the proposed hybrid EfficientNet-Self Attention model has a rapid convergence rate, a high stability rate, and a high generalization rate. These findings ensure the existence of a large gap between training and validation curves and confirm the effectiveness of the attention mechanism in terms of suppressing irrelevant masked areas and boosting identity-discriminative facial features, which is why the 99 percent accuracy in classification is reported and the model can be considered as appropriate in terms of masked face recognition in real-life settings..

Table 1. Performance Metrics

	Precision	**Recall**	**F1-score**	**Support**
Masked_MFR2	0.99	0.98	0.98	332
Unmask_MFR2	0.98	0.99	0.99	342
Accuracy			0.99	674
Macro Avg	0.99	0.99	0.99	674
Weighted Avg	0.99	0.99	0.99	674

The total classification error of the proposed system is 0.99 using 674 total samples which shows almost perfect recognition of the MFR2 dataset as shown in Table 1. The high accuracy confirms the success of the combination of EfficientNet feature extraction and a self-attention mechanism to deal with the facial occlusion.The macroaverage precision, the recall and the F1-score are all found to be equal at 0.99, which means that the model does not favor either masked or unmasked face. The Macro averaging is applied to each class separately and the high values can be attributed to the similarity in performance across the categories.The weighted-average measures which consider the support of classes also provide the values of 0.99 of precision, recall and F1-score. This means that the class imbalance of the dataset does not have an adverse impact on the overall performance of the model.

Table 2. Performance Comparison of Different Model Configurations on MFR2 Dataset

Model Architecture	**Accuracy**	**Precision**	**Recall**	**F1-score**
EfficientNet-B0 (Baseline)	0.94	0.94	0.93	0.93
EfficientNet-B0 + SE Block	0.96	0.96	0.95	0.95
Model Architecture	**Accuracy**	**Precision**	**Recall**	**F1-score**
EfficientNet-B0 + Self-Attention (Proposed)	0.99	0.99	0.99	0.99

In Table 2, comparative assessment of various model settings is provided. The EfficientNet-B0 baseline can already perform competitively, but with the self-attention mechanism included, the performance will significantly increase in all the metrics of evaluation. This shows that attention based feature refinement is effective to suppress mask related feature and strengthen identity based discrimination areas.

Model Complexity and Parameter Analysis

The total number of parameters in the proposed architecture is 4,374,706, as obtained from the model summary output. Among these, 4,330,690 parameters are trainable, indicating that both the EfficientNetV2B0 backbone and the custom classification layers were optimized during training. The remaining 44,016 parameters are nontrainable, primarily corresponding to fixed parameters in Batch Normalization layers such as moving mean and moving variance.

The relatively low parameter count (approximately 4.37 million) demonstrates that the proposed model maintains a favorable balance between representational capacity and computational efficiency. This compact architecture makes it suitable for real-time masked face recognition applications while avoiding excessive memory consumption.

Inference Performance

Although explicit latency profiling was not performed using hardware-specific benchmarking tools, a rough estimation of inference speed was derived from the evaluation logs. The model processed 674 test images across 22 batches in 16 seconds during prediction. This corresponds to an approximate inference time of:

$$0.0237 \text{ seconds per image} \qquad (13)$$

which translates to nearly:

$$42 \text{ images per second} \qquad (14)$$

It is important to note that this measurement includes data loading overhead from the test generator and therefore does not represent pure forward-pass inference time. Nevertheless, the results indicate that the proposed system is computationally efficient and capable of near real-time masked face recognition performance under standard hardware configurations.To enhance generalization and mitigate overfitting, a Dropout layer with a rate of 0.4 was incorporated into the architecture. The dropout operation was applied after the GlobalAveragePooling2D layer in the classification head. During training, 40% of the neurons in this layer are randomly deactivated, reducing co-adaptation among feature detectors and encouraging the model to learn more robust and discriminative representations.

This regularization mechanism is particularly important in masked face recognition tasks, where occlusion caused by facial masks limits the availability of distinctive facial features. By enforcing redundancy in feature learning, dropout improves stability and generalization on unseen masked face samples. During inference, dropout is automatically disabled, allowing the model to utilize its full representational capacity.

Overview of the Mfr2 Dataset

The Real-World masked face recognition dataset (MFR2) is an important element of the experimental design of this study, namely, selected to prove the effectiveness of the developed hybrid model of EfficientNet and selfattention mechanism. In computer vision, especially after the world has changed the standard of facial appearances between 2020 and 2025, high-quality, non-synthetic datasets have been one of the main bottlenecks to creating a strong system of identification. The MFR2 dataset resolves this issue as it provides a set of images taken in uncontrolled real-world scenarios, and hence gives the model a strict test of its adaptability to generalize outside the limited laboratory context. In comparison to datasets obtained with the use of artificial masking, where surgical masks are overlaid onto existing face databases, MFR2 is characterized by the emphasis on the pairs of verification tasks including public figures, which necessitates the use of a model that can be used to address the authentic masking, different light scenarios, or diverse head pose. Based on the available statistical breaks downs, the dataset has 269 images of 53 different identities. This is a rather small size in comparison to large-scale training setups, but is very specialized with regards to benchmarking recognition accuracy in challenging conditions. The images of the dataset are clearly classified to enable cross-condition testing: there are 169 masked face images and 100 non-masked face images. The proposed hybrid model can be assessed with the help of such ratio in the framework of mixed-modality situations, when the system is to match a masked probe image to a non-masked gallery image, or the other way around. The dataset has on average around five images per subject, which is enough to create intra-class variations and also has sufficient number of samples to create challenging sparsity that does not allow the model to overfit in the validation phase.

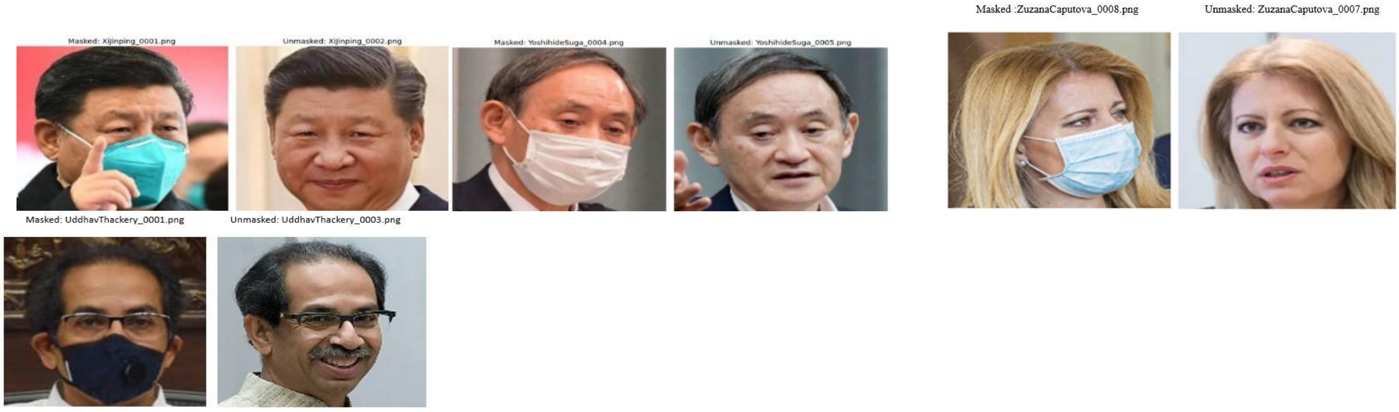
Training and Validation Strategies

To evaluate the proposed model, experiments were conducted on the Masked Face Recognition (MFR2) dataset, a publicly available benchmark specifically curated for the task of recognizing individuals under masked facial occlusion. The dataset comprises real-world face images in which subjects wear various types of face masks, presenting significant challenges to conventional facial recognition systems due to the occlusion of discriminative facial regions such as the nose, mouth, and lower cheeks.

For the purpose of this study, the dataset was partitioned into three distinct subsets following a stratified splitting strategy to ensure a balanced class distribution across all splits. Specifically, 70% of the total available samples were allocated to the training set, providing the model with sufficient labeled examples to learn discriminative feature representations. The remaining samples were divided into a test set comprising 20% of the data, used for final performance evaluation, and a validation set comprising 10% of the data, employed during training to monitor generalization performance and guide early stopping and learning rate scheduling decisions.

The proposed architecture, consisting of a fine-tuned EfficientNetV2B0 backbone augmented with a custom self-attention mechanism, was trained on the MFR2 training partition for a total of five epochs. Although the number of training epochs is relatively modest, the combination of transfer learning from ImageNet pre-trained weights and the adaptive optimization strategy employed ensured that meaningful convergence was achieved within this limited training duration. The self-attention mechanism further enhanced the model's ability to focus on unoccluded discriminative facial regions, such as the eyes and forehead, compensating for the loss of information introduced by mask occlusion and enabling robust identity recognition under challenging real-world conditions.

Figure 5. Real-World Image of MFR2 Dataset


Masked: XiJinping_0001.png
Unmasked: XiJinping_0002.png
Masked: YoshihideSuga_0004.png
Unmasked: YoshihideSuga_0005.png
Masked :ZuzanaCaputova_0008.png
Unmasked: ZuzanaCaputova_0007.png
Masked: UddhavThackery_0001.png
Unmasked: UddhavThackery_0003.png

The MFR2 data set gives the required empirical basis of evaluating the recommended hybrid approach. Its statistical structure 53 identities, 269 overall images, and 169 masked and 100 non-masked samples is a brief, but still impressive challenge that recreates the situation of real-world surveillance and media observation. With the help of this dataset, the study is no longer an abstract architecture-design experiment but a practical one, to guarantee that the high success rates obtained can be transferred to the real challenges of detecting individuals wearing a mask in the field.

CONCLUSION

This study proposed a powerful hybrid deep learning architecture of masked face recognition by combining EfficientNet-B0 with a self-attention mechanism. The suggested method can also be seen as an effective solution to the difficulties of facial occlusion by focusing on identity-saving areas of the face and inhibiting features of the mask. EfficientNet-B0 has efficient and discriminative feature extraction and the self-attention module leads to high recognition accuracy of 99 percent, and the values of precision, recall, and F1-score are high and stable in both masked and unmasked face classes. The stability of the model and the possibility to generalize is also evidenced by the close correspondence between the training and the validation performance. The effectiveness of attention-based feature refinement in alleviating the effect of occlusion is justified by the classification report and learning curves.

The experimental results also confirm that the proposed algorithm is effective in identifying individual faces in masked face images, and can therefore be used in real-world biometric systems like access control, surveillance and identity verification systems. Future research can consider how to incorporate lightweight transformer architectures, testing over larger, more diverse datasets, and how to be resource-efficient to run on resourceconstrained edge devices.

REFERENCES

1. W. Wan et al., "Masked face recognition via dual-branch convolutional self-attention network," Applied Soft Computing, 2024, doi: 10.1016/j.asoc.2024.112595.
2. A. F. et al., "Ensemble Learning using Transformers and Convolutional Networks for Masked Face Recognition," 2022, doi: 10.1109/sitis57111.2022.00070.
3. I. Hosen et al., "HiMFR: A hybrid masked face recognition through face inpainting," arXiv.org, 2022, doi: 10.48550/arXiv.2209.08930.
4. Mohammed R. Al-Sinan et al., "Ensemble Learning using Transformers and Convolutional Networks for Masked Face Recognition," in International Conference on Signal-Image Technology and Internet-Based Systems, 2022, doi: 10.1109/SITIS57111.2022.00070.
5. Z. Wang, G. Wang, B. Huang, Z. Xiong, Q. Hong, H. Wu, P. Yi, K. Jiang, N. Wang, Y. Pei, et al. Masked face recognition dataset and application. arXiv preprint arXiv:2003.09093, 2020.
6. Mohammed R. Al-Sinan et al., "Ensemble Learning using Transformers and Convolutional Networks for Masked Face Recognition," 2022, doi: 10.48550/arxiv.2210.04816.
7. M. Rodrigo et al., "Comprehensive comparison between vision transformers and convolutional neural networks for face recognition tasks," Dental Science Reports, 2024, doi: 10.1038/s41598-024-72254-w.
8. I. Hosen et al., "HiMFR: A Hybrid Masked Face Recognition Through Face Inpainting," 2022.
9. Y. Ge et al., "Masked face recognition with convolutional visual self-attention network," Neurocomputing, 2022, doi: 10.1016/j.neucom.2022.10.025.
10. W. Zhao et al., "Masked Face Transformer," IEEE Transactions on Information Forensics and Security, doi: 10.1109/tifs.2023.3322600.
11. Y. Zhu et al., "Joint holistic and masked face recognition," IEEE Transactions on Information Forensics and Security, doi: 10.1109/TIFS.2023.3280717.
12. W. Wan et al., "Masked face recognition based on knowledge distillation and convolutional self-attention network," International Journal of Machine Learning and Cybernetics, 2024, doi: 10.1007/s13042-02402390-2.
13. L. Ouannes et al., "Enhancing Face Recognition in Degraded Conditions via Vision Transformer," 2024, doi: 10.1109/codit62066.2024.10708244.
14. L. C. Tiong et al., "Flexible Biometrics Recognition: Bridging the Multimodality Gap Through Attention, Alignment and Prompt Tuning," 2024, doi: 10.1109/cvpr52733.2024.00033.
15. Y. Wang et al., "Learning 3d face representation with vision transformer for masked face recognition," 2022, doi: 10.1109/CACML55074.2022.00092.
16. K. Wang et al., "FaceMAE: Privacy-Preserving Face Recognition via Masked Autoencoders," 2022.
17. J. Guo et al., "Face Recognition System with Occlusion Based on Attention Mechanism Improvement of the Vision Mamba Model," 2025, doi: 10.1109/icaisisas64483.2025.11051738.
18. H. V. Phan et al., "Fast and interpretable face identification for out-of-distribution data using vision transformers," arXiv.org, 2023, doi: 10.48550/arxiv.2311.02803.
19. A. Zhu et al., "Cross-Task Multi-Branch Vision Transformer for Facial Expression and Mask Wearing Classification," 2024, doi: 10.48550/arxiv.2404.14606.
20. W. Chang et al., "ResSaNet: A Hybrid Backbone of Residual Block and Self-Attention Module for Masked Face Recognition," in International Conference on Computer Vision, 2021, doi: 10.1109/ICCVW54120.2021.00170.
21. S. Lee et al., "Latent-OFER: Detect, Mask, and Reconstruct with Latent Vectors for Occluded Facial Expression Recognition," 2023, doi: 10.1109/iccv51070.2023.00148.
22. V. Vidal et al., "A benchmark on masked face recognition," in SIBGRAPI Conference on Graphics, Patterns and Images, 2022, doi: 10.1109/SIBGRAPI55357.2022.9991785.
23. Y. Xu, "Based on the contrastive learning classifier for occluded face recognition," Procedia Computer Science, 2025, doi: 10.1016/j.procs.2025.08.148.
24. Y. Liu, "Masked face recognition based on transfer learning."
25. A. Iftikhar et al., "Masked Face Detection and Recognition Using a Unified Feature Extractor," 2024, doi: 10.1109/icacs60934.2024.10473243.

26. S. Yang et al., "Hybrid Architecture-Based Evolutionary Robust Neural Architecture Search," IEEE Transactions on Emerging Topics in Computational Intelligence, 2024, doi: 10.1109/tetci.2024.3400867.
27. Q. Wang et al., "AAN-Face: Attention Augmented Networks for Face Recognition," IEEE Transactions on Image Processing, 2021, doi: 10.1109/TIP.2021.3107238.
28. H. Qiu et al., "End2End Occluded Face Recognition by Masking Corrupted Features," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, doi: 10.1109/TPAMI.2021.3098962.
29. Q. Wang et al., "DSA-Face: Diverse and Sparse Attentions for Face Recognition Robust to Pose Variation and Occlusion," IEEE Transactions on Information Forensics and Security, 2021, doi: 10.1109/TIFS.2021.3109463.
30. A. George et al., "EdgeFace: Efficient Face Recognition Model for Edge Devices," 2023.