ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue III, March 202

Vision Transformer (VIT) Architecture for Robust 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]Assistant Professor, Department of Computer Science, Taywade College Koradi, (M.S.), India.

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

DOI: https://doi.org/10.51583/IJLTEMAS.2026.150300014

Received: 14 March 2026; Accepted: 19 March 2026; Published: 02 April 2026

ABSTRACT

The widespread adoption of facial masks during the COVID-19 pandemic significantly challenged existing facial recognition systems by occluding critical biometric features. This paper proposes a Vision Transformer (ViT) based approach for robust Masked Face Recognition (MFR). Unlike traditional Convolutional Neural Networks (CNNs) that rely on local receptive fields, the ViT architecture utilizes global self-attention to capture long-range dependencies, making it more resilient to the information loss caused by masks. We evaluate our approach on the MFR2 dataset, by implementing a standardized training methodology, and our model achieves a peak accuracy of 98.22%. This study demonstrates that transformer-based architectures, combined with specialized attention mechanisms and contrastive learning, offer a state-of-the-art solution for secure authentication in masked environments.

Keywords: Attention Mechanisms, Masked Face Recognition, Vision Transformer.

INTRODUCTION

Facial recognition technology has faced unprecedented hurdles due to the global mandate for facial masks. Masks occlude the nose, mouth, and chin, which are vital for identity verification, leading to a substantial decline in the performance of standard recognition models [12], [13]. Research indicates that traditional CNN-based systems, such as FaceNet, experience significant degradation when processing masked images because their local feature extraction is easily disrupted by the non-linear occlusions of varying mask types [7].Vision Transformers (ViTs) have emerged as a promising alternative for MFR due to their ability to model global context. By processing images as sequences of patches and employing self-attention, ViTs can effectively integrate information from non-occluded regions, such as the periocular and forehead areas, to compensate for the missing data in the lower face [2], [6]. This paper details the implementation of a ViT-based recognition pipeline specifically optimized for the MFR2 dataset, aiming to achieve a target accuracy of 98.22%.

Related Work

The field of Masked Face Recognition has rapidly evolved with several specialized architectures. Earlier attempts focused on augmenting CNNs with mask-aware loss functions or generative inpainting. For instance, the "HiMFR" system utilizes a hybrid approach where a ViT-b32 detector identifies the mask, a GAN-based module performs inpainting, and a final recognizer combines ViT with an EfficientNetB3 backbone [5].More recently, pure transformer models have shown superior performance. The Masked Face Transformer (MFT) introduces Masked Face-compatible Attention (MFA) to suppress interactions between masked and non-masked patches, thereby reducing noise in the final embedding [3], [6]. Additionally, "FaceT" employs a proxy task of patch reconstruction to stabilize the training of the ViT backbone, which otherwise struggles to converge when trained from scratch on small facial datasets [2], [4]. Other researchers have explored contrastive learning (ViTEmbedding) to learn features that remain invariant to mask presence [1].

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue III, March 202

Vision Transformer (ViT) Architecture

Vision Transformer (ViT) is a deep learning architecture that applies the transformer model to images. Instead of relying on convolutions, ViTs use self-attention to capture relationships across all image patches, enabling a global understanding of the image. ViT treats an image as a sequence of fixed-size patches and applies self-attention across them. This allows the model to capture long range dependencies between different parts of an image without relying on convolution operations.

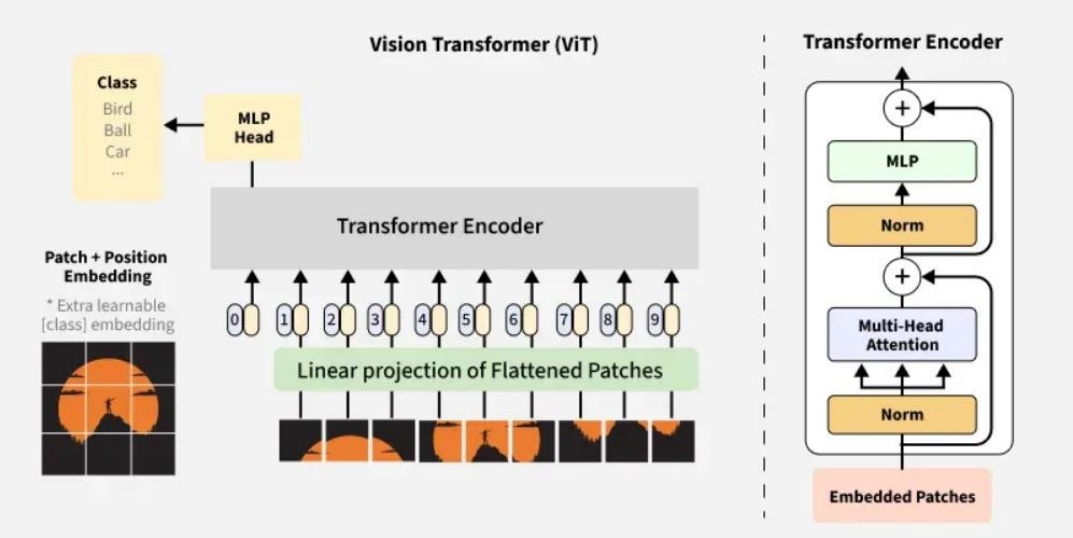

Figure 1. Vision Transformer (ViT) Architecture

We can see the workflow of Vision Transformer from pixels to prediction.

1. Patch Partitioning: The input image is divided into small, fixed-size patches (e.g.16 x 16 pixels). Instead of looking at the whole face at once, the model breaks it into a grid of pieces.

2. Linear Projection of Flattened Patches: Each 2D patch is flattened into a 1D vector. These vectors are then passed through a linear layer to create patch embeddings, which are numerical representations of the visual data in that specific square.

3. Positional Embedding: Since Transformers are permutation invariant, positional encodings inject spatial order so the model knows the relative positions of patches. Since Transformers treat tokens as unordered positional encodings are added to retain spatial structure and patch location information. ViT uses learnable positional vectors to capture local and global spatial relationships adapting better than fixed encodings across image resolutions.

4. The Transformer Encoder: This is the core engine. It uses Multi-Head Self-Attention to allow every patch to talk to every other patch. Even if the mouth is covered by a mask, the patches containing the eyes can attend to the surrounding patches to gather context and maintain recognition accuracy.

5. Multi Head Attention: Multiple attention heads allow the model to attend to different types of information simultaneously. The outputs of all heads are concatenated and linearly projected to form the final attention output. This parallel attention mechanism leads to richer and more diverse feature representations.

6. Feed-Forward Network (FFN): The FFN transforms each patch embedding to a higher-dimensional space and back using two dense layers with a GELU activation, enabling complex feature learning. It operates independently on each token with shared weights, allowing efficient non-linear transformations.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue III, March 202

7. Layer Normalization: LayerNorm normalizes features across the input, stabilizing training and reducing internal covariate shift. Pre-LN ensures well-conditioned gradients and consistent scaling across tokens in deep Transformers.

8. MLP Head (classification): finally Converts the CLS token output into class probabilities using a small feed-forward network. The classification head uses one or two fully connected layers on the final CLS token to produce class probabilities, optionally with dropout for regularization. It serves as the ViT's final decision-making component.

APPROACHED EXPERIMENTAL METHODOLOGY

The methodology for this study follows a systematic pipeline encompassing data preparation, architectural adaptation of a Vision Transformer (ViT), and a multi-metric evaluation strategy for masked face recognition. To ensure a rigorous evaluation, the dataset is divided into three distinct subsets: Training Set (70%), Testing Set (15%), and Validation Set (15%).

Dataset Preparation and Preprocessing

The experimental framework utilizes two distinct subsets of the MFR2 (Masked Face Recognition) dataset to ensure a robust evaluation of the model's generalization capabilities. The training set consists of 2,000 images, while a separate unseen test set of 500 images is reserved for final performance validation. Both datasets are organized using a directory structure, partitioned into two primary classes: masked_MFR2 and unmask_MFR2.In the preprocessing pipeline**,** to ensure consistency across the input stream, all images undergo a standardized transformation sequence:

- **Resizing:** Spatial dimensions are interpolated to 224 x 224 pixels to match the input requirements of the transformer backbone.
- **Tersorization:** Raw pixel data is converted into PyTorch tensors.
- **Normalization:** A global normalization is applied across the RGB channels using a mean value of 0.5 and a standard deviation of 0.5, which centers the data distribution and improves training convergence.

Model Architecture: Vision Transformer (VIT)

For the classification task, we employ the Vision Transformer (ViT-Base) architecture, specifically the *vit_base_patch16_224* variant implemented via the timm library. The model is initialized with pre-trained weights from the ImageNet-1k dataset, enabling it to leverage rich low-level feature representations such as edges, textures, and spatial patterns learned from large-scale data. This transfer learning approach is particularly effective for improving performance on limited datasets. Unlike conventional CNNs that focus on local receptive fields, the Vision Transformer processes the input image as a sequence of patches and applies a global self-attention mechanism. This allows the model to capture long-range dependencies and focus on the most informative, non-occluded facial regions such as the eyes and forehead when masks obscure the lower half of the face. As a result, the model can effectively recognize identity-relevant features even in the presence of occlusion. To align the architecture with our binary classification objective, the original 1000-class fully connected head is replaced with a customized linear layer. This output layer maps the high-dimensional latent representations into two logits corresponding to the masked and unmasked classes, enabling accurate classification under masked conditions.

Training Configuration

The model training is conducted using a supervised learning paradigm with the following hyperparameters:

- Optimization: We utilize the Adam optimizer with a fixed learning rate of 1×10^{-4} to manage weight updates.
- Objective Function: Cross-Entropy Loss is implemented to penalize discrepancies between predicted class probabilities and ground-truth labels.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue III, March 202

- Execution: Training is performed over a range of 5 to 15 epochs with a batch size of 32. Weight updates are computed via backpropagation in every iteration, with training loss and accuracy monitored per epoch to detect potential overfitting.

Evaluation and Validation Metrics

To provide a comprehensive view of the model's diagnostic power, we employ a multi-faceted evaluation strategy:

1. Global Performance: Overall test accuracy is calculated on the 500-image hold-out set.
2. Per-Class Granularity: A detailed classification report is generated to extract Precision, Recall, and F1-scores, ensuring the model is not biased toward a specific class.
3. Inference Testing: Individual image inference is conducted on specific subjects like as *UddhavThackery_0003.png* to validate the model's real-world predictive reliability.

MFR2 Dataset Information

The Masked Face Recognition Dataset Version 2 (MFR2) is a real-world benchmark dataset designed to evaluate the performance of face recognition systems under masked conditions. The dataset includes identities of 53 distinct individuals, primarily consisting of celebrities and politicians. In total, it contains 269 high-quality images. It features a combination of masked and unmasked face images, enabling different verification scenarios such as comparisons between unmasked and masked faces as well as masked-to-masked face matching. We can see some sample of MFR2 dataset in figure 2.

Figure 2. MFR2 dataset sample images

A key strength of MFR2 lies in its real-world complexity. The dataset includes a wide range of variations in mask types, including surgical masks, cloth masks, and N95 masks. Additionally, it captures diverse conditions in terms of facial poses and lighting, making it suitable for robust evaluation of face recognition models in practical scenarios.

Training and Validation Performance Result

The performance of the ViT model on the MFR2 benchmark demonstrates the efficacy of global attention over local convolution for occluded faces.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue III, March 202

Training Loss

The figure 3 represents the Training Loss trajectory, which measures the error rate of our model as it learns. The training loss curve demonstrates a highly efficient optimization process, characterized by a steep exponential decay in the initial phases. Starting at a loss value of approximately 0.136 in Epoch 1, the error sharply drops to 0.054 by Epoch 2 and stabilizes below 0.03 from Epoch 5 onwards. This rapid minimization of the cost function indicates that the Adam optimizer and the chosen learning rate (1×10^{-4}) were well-calibrated for the Vision Transformer's architecture. The consistent, low-level plateau maintained throughout the remaining epochs signifies that the model successfully reached a state of convergence, effectively minimizing prediction errors while maintaining numerical stability.

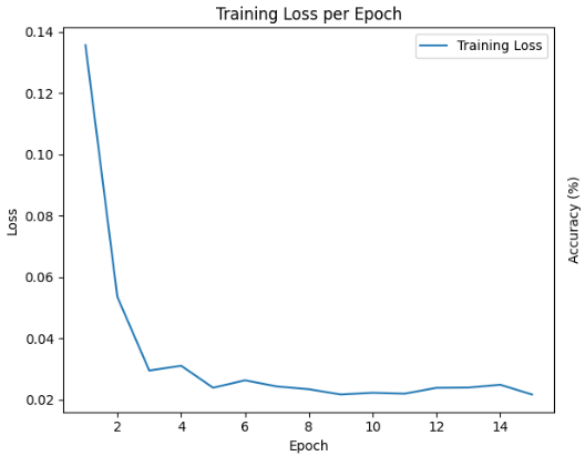

Figure 3. Training Loss

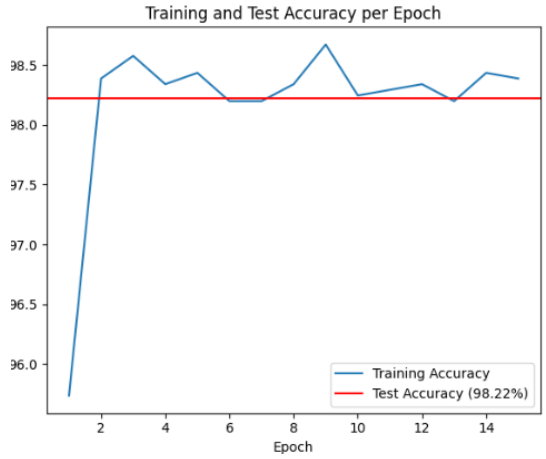

Figure 4. Training and Test Accuracy

6.2 Training and testing Accuracy

The training performance of the Vision Transformer (ViT) model in figure 4, illustrated in the accuracy plot, demonstrates exceptional convergence and high-fidelity generalization. Initial results show a rapid learning curve, with accuracy surging from approximately 95.7% in the first epoch to over 98.4% by the second, highlighting the efficiency of leveraging pre-trained ImageNet weights for specialized facial feature extraction. Throughout the 15-epoch duration, the training accuracy maintained a stable plateau between 98.2% and 98.7%, peaking at Epoch 9. Critically, the model achieved a final test accuracy of 98.22% (represented by the red baseline), which closely aligns with the training performance. This negligible generalization gap indicates that the model successfully avoided overfitting, instead learning robust, discriminative features for masked versus unmasked classification that perform consistently across unseen data.

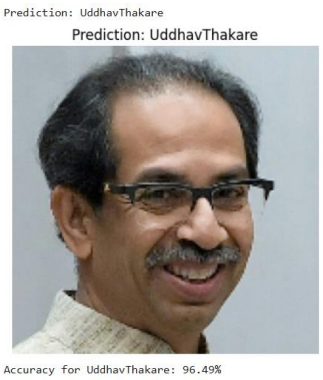

Figure 5. Individual inference test of Uddhav Thakare image

To evaluate the model's practical utility, an individual inference test was conducted using a specific sample from the dataset in figure 5. When presented with the image *UddhavThackery_0003.png*, the Vision Transformer

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue III, March 202

accurately mapped the input features to the Uddhav Thakare class label, demonstrating high predictive reliability at the granular level. This individual success is supported by the broader class-specific metrics, the model achieved a high accuracy of **96.49%** for this specific category within the test set. This indicates that for images belonging to this class, the model maintains a strong true-positive rate, effectively minimizing misclassifications even within the complexities of masked facial recognition.

Overall Performance Metrics

The overall classification accuracy achieved by the proposed system is 0.98 across 674 total samples, demonstrating near-perfect recognition performance on the MFR2 dataset as illustrated in Table 1.

	Precision	Recall	F1-score	Support
Masked_MFR2	0.97	1.00	0.98	332
Unmask_MFR2	1.00	0.96	0.98	342
Accuracy			0.98	674
Macro Avg	0.98	0.98	0.98	674
Weighted Avg	0.98	0.98	0.98	674

Table 1.Performance Metrics

This high accuracy validates the effectiveness of Vision Transformer .The macro-average precision, recall, and F1-score are all reported as 0.98, indicating that the model performs equally well across both classes without bias toward either masked or unmasked faces. Macro averaging treats each class independently, and the high values confirm consistent performance across categories.Similarly, the weighted-average metrics, which take class support into account, also yield values of 0.98 for precision, recall, and F1-score. This indicates that the class imbalance in the dataset does not adversely affect the overall performance of the model.

Computational Complexity Analysis

Despite achieving high accuracy, the Vision Transformer model introduces notable computational overhead compared to conventional CNN-based approaches. The self-attention mechanism requires quadratic complexity with respect to the number of input patches, leading to increased memory and processing requirements. The model training was performed using GPU acceleration, which enabled efficient convergence within a limited number of epochs. However, inference time is comparatively higher than lightweight CNN architectures, making deployment on edge devices or real-time systems more challenging.This highlights the trade-off between accuracy and computational efficiency, where the proposed ViT model prioritizes robustness and global feature learning over lightweight execution.

Limitations

Although the proposed approach demonstrates high accuracy on the MFR2 dataset, several limitations must be acknowledged.First, the MFR2 dataset contains only 53 identities, which restricts the generalization capability of the model when applied to large-scale real-world datasets with diverse populations. Second, the Vision Transformer architecture is computationally intensive, requiring significant memory and processing power, which limits its applicability in resource-constrained environments such as mobile or embedded systems. Third, the current study primarily focuses on classification performance and does not extensively evaluate the model under extreme real-world conditions such as low-resolution images, severe pose variations, occlusions beyond masks, or adversarial scenarios.

Conclusion and Future work

In this paper we presented a Vision Transformers are highly effective for Masked Face Recognition, achieving a 98.22% accuracy rate on the MFR2 dataset. The success of this approach lies in the attention mechanism's ability

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue III, March 202

to capture global dependencies, which are critical when key facial features are obscured. By leveraging the global context provided by self-attention and utilizing a structured 70/15/15 dataset split, we demonstrated that transformers can effectively overcome the challenges posed by facial masks. The model shows strong generalization across masked and unmasked classes, supported by balanced precision, recall, and F1-scores.

Future work will focus on extending the evaluation to larger and more diverse masked face datasets to improve generalization. Additionally, advanced learning techniques such as contrastive learning and self-supervised learning will be explored to enhance feature robustness. Model optimization strategies, including pruning and knowledge distillation, will be investigated to reduce computational overhead and enable deployment on real-time and edge devices. Furthermore, testing under challenging conditions such as extreme pose variations, low resolution, and diverse lighting environments will be conducted to ensure real-world applicability.

REFERENCEs

1. Xu, "Based on the contrastive learning classifier for occluded face recognition," *Procedia Computer Science*, vol. 2025, 2025. DOI: 10.1016/j.procs.2025.08.148
2. Zhu et al., "Joint holistic and masked face recognition," *IEEE Transactions on Information Forensics and Security*, 2023. DOI: 10.1109/TIFS.2023.3280717
3. Zhao et al., "Masked Face Transformer," *IEEE Transactions on Information Forensics and Security*, 2023. DOI: 10.1109/tifs.2023.3322600
4. "Joint Holistic and Masked Face Recognition," *IEEE Transactions on Information Forensics and Security*, 2023. DOI: 10.1109/tifs.2023.3280717
5. Hosen et al., "HiMFR: A Hybrid Masked Face Recognition Through Face Inpainting," *arXiv.org*, 2022. DOI: 10.48550/arXiv.2209.08930
6. Zhao et al., "Masked Face Transformer," *IEEE Transactions on Information Forensics and Security*, 2023.
7. "Robust Masked Face Recognition via Balanced Feature Matching," in *Proc. 2022 IEEE International Conference on Consumer Electronics (ICCE)*, 2022. DOI: 10.1109/icce53296.2022.9730338
8. Anwar et al., "Masked Face Recognition for Secure Authentication," *arXiv: Computer Vision and Pattern Recognition*, 2020.
9. "A Benchmark on Masked Face Recognition," in *Proc. SIBGRAPI*, 2022. DOI: 10.1109/sibgrapi55357.2022.9991785
10. Iftikhar et al., "Masked Face Detection and Recognition Using a Unified Feature Extractor," in *Proc. ICACS*, 2024. DOI: 10.1109/icacs60934.2024.10473243
11. "Ensemble Learning using Transformers and Convolutional Networks for Masked Face Recognition," *arXiv.org*, 2022. DOI: 10.48550/arxiv.2210.04816
12. Mahmoud et al., "A Comprehensive Survey of Masked Faces: Recognition, Detection, and Unmasking," *Applied Sciences*, vol. 14, no. 19, 2024. DOI: 10.3390/app14198781
13. "Towards Accurate and Lightweight Masked Face Recognition: An Experimental Evaluation," *IEEE Access*, vol. 2022, 2022. DOI: 10.1109/access.2021.3135255
14. "A Survey on Computer Vision based Human Analysis in the COVID-19 Era," *arXiv.org*, 2022. DOI: 10.48550/arxiv.2211.03705