Expert Systems

WILEY

Expert Systems

ORIGINAL ARTICLE

Spatial-Frequency Feature Fusion Based Deepfake Detection With Mask Supervision

Yilin Zhang[1] | Zhuocheng Wu[2] | Yufeng Song[3] | Fei Wang[3] | Zengren Song[4] | Bo Wang[3]

[1]School of Future Technology, Dalian University of Technology, Dalian, Liaoning, China | [2]School of Computer Science and Technology, Dalian University of Technology, Dalian, Liaoning, China | [3]School of Information and Communication Engineering, Dalian University of Technology, Dalian, Liaoning, China | [4]National Computer Network Emergency Response Technical Team, Coordination Center of China, Beijing, China

Correspondence: Bo Wang (bowang@dlut.edu.cn)

Received: 19 April 2025 | **Revised:** 13 November 2025 | **Accepted:** 19 January 2026

Keywords: compressed images | deepfake detection | mask supervision | multi-modal feature fusion

ABSTRACT

In recent years, artificial intelligence technologies have been widely used, bringing security, convenience and certain risks. Deepfake techniques raise significant security concerns by manipulating facial images to create convincing but false representations. We introduce a novel mask-supervision-based deepfake detection method to improve detection performance in this context. Our approach corrects the model's focus on irrelevant regions through mask supervision, using pixel-level labels to guide the model towards synthetic facial regions and ensure more accurate extraction of spatial features. In addition, we incorporate a frequency-domain feature extraction module that exploits the robustness of frequency-domain cues to compression artefacts. We first preprocess the input image and then feed it into the mask supervision and frequency-domain feature extraction modules. The mask supervision module extracts intermediate features using the High-Resolution Network (HRNet) and refines spatial features by guiding the prediction mask with the ground-truth mask. The frequency-domain module extracts features via the Discrete Cosine Transform (DCT) and filtering across different frequency bands. Finally, spatial and frequency-domain features are concatenated and fed into a classification network to output the final prediction. Experimental results show that our method maintains good robustness in compressed scenarios.

1 | Introduction

Deepfake is an artificial intelligence-based synthesis and editing technology for multimedia content such as images (Keita et al. 2025), audio (Dixit et al. 2023) and video (Ramadhani et al. 2024). Malicious use of deepfake poses a significant threat to personal information security and social stability (Park et al. 2024). For example, deepfake fraud is a growing concern. In January 2024, John Lee, the Chief Executive of the Hong Kong Special Administrative Region, was targeted by a deepfake attack—an online fake video falsely claimed he endorsed a high-return investment scheme. The Hong Kong government promptly clarified the situation and confirmed the video contained false information. Deepfake pornography is another major malicious application of the technology. Software like Clothoff and Nudify utilises deepfake technology to perform 'one-click stripping'. According to Wire magazine, thousands of people use DeepNude monthly to create nude photos of friends and family, including underage victims. Despite national policies restricting DeepNude's use, the spread of its content remains hard to curb due to the complexity and diversity of social platform data (Khder et al. 2023). Deepfake detection has advanced significantly in recent years, with continuous innovative research (Passos et al. 2024). Detection methods have been developed for spatial, frequency and cross-modal domains (Sudarshana and Vamsidhar 2025), whereas interdisciplinary

Abbreviations: DCT, discrete cosine transform; FF++, FaceForensics++; GAN, Generative Adversarial Network; HRNet, High-Resolution Network.

approaches—such as semantic segmentation, model compression and forged speech detection—have injected new momentum into the field (Jannu et al. 2024). Despite outperforming traditional machine learning methods, deepfake detection in compressed scenarios remains a critical challenge. Most images and videos shared on social platforms undergo post-processing like compression and filtering, which weaken or even eliminate detection cues. This leads to significant performance degradation or complete failure of existing methods. Recent studies have explored solutions to enhance generalisation: cross-modal alignment and distillation (Du et al. 2024), meta-learning-based efficient vision transformers (Tran et al. 2024), and forgery-aware audio-distilled learning (Nie et al. 2024). For instance, Liu et al. (2024) achieved 99.9% accuracy on the FaceForensics++ (FF++) dataset by exploiting inconsistent expression cues, but the method's robustness to compressed content was not explicitly verified. Similarly, methods relying on image decomposition (Lai et al. 2024) or mask-guided supervision (Li et al. 2024) focus on spatial-domain features but lack integration with frequency-domain cues that resist compression artefacts. We attribute the significant performance drop of existing models on low-quality data to compression's ability to erase or weaken forensically relevant artefacts. To address this, we design two core strategies: first, integrate spatial and frequency-domain information to capture more comprehensive features; second, reduce the loss of key features during compression by redirecting the network's attention to artefact-rich regions. To improve detection performance on compressed facial images, we propose a mask-supervision-based deepfake detection method. Leveraging spatial-frequency features' resistance to compression, we first design a frequency-domain feature extraction module to capture image frequency information. We then introduce a mask supervision mechanism during spatial feature extraction—using pixel-level labels to guide the model towards forged regions, enabling more accurate focus on altered parts. Finally, we concatenate spatial and frequency-domain features and feed them into a classification network to generate final predictions. Our contributions are summarised as follows:

- We propose a mask-supervision-based deepfake detection method that fuses spatial and frequency-domain features, tailored for forged images with varying compression levels. This integration mitigates compression's impact on detection effectiveness.
- We design a complete detection framework encompassing preprocessing, mask supervision, frequency-domain feature extraction and classification modules. Their synergistic operation enhances feature expressiveness while improving detection accuracy and robustness.
- We conduct extensive experiments on public datasets with different compression levels, fully validating the proposed method. Results demonstrate significant performance advantages in detecting compressed forged images.

This paper is structured as follows: Section 2 reviews related work on deepfake detection, categorising existing methods by feature type and identifying critical research gaps. Section 3 details the proposed methodology, including the overall framework, module designs and loss function formulation. Section 4 presents comprehensive experimental results, including performance evaluations across uncompressed and compressed scenarios, ablation studies to verify component effectiveness, and visualisations to interpret model behaviour. Finally, Section 5 summarises the study's conclusions, discusses limitations and outlines directions for future research.

2 | Related Work

Deepfake detection methods are categorised by extracted feature types into hand-crafted feature-based and deep feature-based approaches. Deep feature-based methods have gradually become mainstream, and we further classify their extracted features into three categories: spatial-domain, frequency-domain and spatial-frequency fusion features—aligning with the core focus of this study.

2.1 | Spatial-Domain Deepfake Detection

The spatial domain (or image space) focuses on pixel-level information, with typical features including biometric, texture/edge, identity and latent features. Recent advances in spatial-domain methods have emphasised supervision mechanisms and generalisation enhancements, but gaps remain in robustness to compression and targeted focus on forged regions.

2.1.1 | Biometric Features

Biometric cues leverage physiological or behavioural characteristics that are difficult to forge. Nguyen and Derakhshani (2020) validated eyebrows as effective indicators, as their edge-of-face location often retains fusion boundaries during face swapping. Agarwal et al. (2020) combined lip movements with phoneme correspondence (e.g., M, B, P) for multimodal detection, but this relies on paired audio-visual data and fails when speech is absent or distorted. A notable recent work by Liu et al. (2024) explored inconsistent facial expressions as biometric-related cues, achieving 99.9% accuracy on the FaceForensics++ (FF++) dataset. However, this method does not validate robustness to compressed content or cross-dataset generalisation (e.g., Celeb-DF2), limiting its practical applicability.

2.1.2 | Texture and Edge Features

Fusion artefacts (e.g., forged boundaries, GAN fingerprints) are key spatial cues for detection. Sun et al. (2021) used facial landmark sequences and recurrent networks to identify unnatural expressions and discontinuous movements. Zhao et al. (2021) enhanced texture distinguishability via attention modules, whereas data augmentation strategies have advanced to isolate artefact features: Li et al.'s 'Facial X-ray' (Li et al. 2020) generates fake faces to highlight fusion boundaries, and Shiohara and Yamasaki (2022) proposed 'self-face-swapping' to exclude irrelevant identity information. Lei et al. (2024) integrated multi-feature decision fusion to defend against adversarial attacks, but the method's reliance on

high-quality texture details leads to performance degradation on compressed images—a critical limitation for real-world social media data.

2.1.3 | Identity Features

Identity consistency is a core anchor for deepfake detection, as forgeries often involve mismatched identities. Dong et al. proposed two identity-based methods: one comparing peripheral facial regions after removing mid-region information (Dong et al. 2020), and another calculating feature vector differences between central and peripheral face areas (Dong et al. 2022). Cozzolino et al. (2021) introduced 3DMM-based identity extraction and temporal verification modules, but these methods struggle with identity-preserving forgeries (e.g., expression manipulation) or low-resolution data. Alshehri (2025) adopted supervised contrastive learning to enhance identity feature scalability and interpretability, though this work also lacks validation on compressed datasets.

2.1.4 | Latent Features and Mask Supervision

Latent features—extracted via backbones like MesoNet (Afchar et al. 2018), Xception (Chollet 2017), ResNet (He et al. 2016) and EfficientNet (Tan and Le 2021)—offer strong generalisability but are sensitive to post-processing. Recent work has integrated mask supervision to guide the model towards forged regions: Li et al. (2024) proposed mask-guided supervision with domain generalisation, using pixel-level labels to focus on tampered areas. Although this improves target region alignment, the method does not integrate frequency-domain cues that resist compression, leaving it vulnerable to real-world post-processing operations.

2.2 | Frequency-Domain Deepfake Detection

Frequency-domain methods address the fragility of spatial features by leveraging cues robust to compression, filtering and resolution degradation. Common transformations include wavelet transform and Discrete Cosine Transform (DCT), with recent advances focusing on anisotropic features and GAN fingerprint extraction. Jia et al. (2021) trained a dual-branch network on static wavelet decomposition to capture intra-image inconsistencies. Wolter et al. (2022) compared wavelet coefficients of real and fake images, retaining partial spatial information for detection. DCT—closely linked to JPEG compression—has been widely adopted: Qian et al.'s F3-Net (Qian et al. 2020) used DCT to extract complementary frequency-aware cues for low-quality fake face detection, and Frank et al. (2020) employed 2D DCT for efficient frequency-domain feature extraction. For GAN fingerprint detection, Jeong et al. (2022) designed FrePGAN to generate frequency-level perturbation maps, enhancing detector generalisation. Huang et al. (2022) exploited anisotropic convolution artefacts in GAN-generated content via fully separable wavelet transforms. A recent extension by Lai et al. (2024) integrated image decomposition with advanced lighting information analysis in the frequency domain, improving robustness to illumination variations. However, most frequency-domain methods lack targeted supervision (e.g., mask guidance) to focus on forged regions, limiting their ability to distinguish subtle artefacts in compressed scenarios.

2.3 | Spatial-Frequency Fusion Deepfake Detection

To capture comprehensive features, recent research has integrated spatial and frequency-domain cues—though gaps remain in combining fusion with targeted supervision and validating compression robustness. Early fusion methods laid the foundation: Liu et al. (2021) proposed a shallow network fusing spatial images and phase spectra to detect upsampling artefacts; Li et al. (2021) designed a spatial-frequency fusion module with single-center loss to enhance feature separability; Luo et al. (2021) integrated SRM-extracted high-frequency features with spatial features, achieving strong cross-dataset performance; Gu et al. (2022) developed a dual-branch network for fine-grained spatial and frequency feature extraction; and Wang et al. (2023) combined wavelet-transformed features with spatial features for classification. Recent state-of-the-art (SOTA) fusion methods have expanded to multi-modal and meta-learning paradigms: Tran et al.'s MEViT (Tran et al. 2025) used meta-learning with EfficientNet-Vision Transformer (ViT) to improve generalisation across forgery types; Du et al.'s CAD framework (Du et al. 2024) integrated cross-modal alignment and distillation for video deepfake detection; and Nie et al.'s FRADE (Nie et al. 2024) adopted forgery-aware audio distillation for multimodal learning. Although these methods advance generalisation and multi-modal integration, they either lack mask supervision to focus on forged regions (e.g., (Tran et al. 2025), (Du et al. 2024)), or do not explicitly validate performance on compressed data (e.g., (Nie et al. 2024), (Liu et al. 2024)).

2.4 | Summary of Research Gaps

Existing methods have made significant progress in spatial supervision, frequency robustness and feature fusion, but three critical gaps remain—directly addressing reviewer concerns: Compression Robustness: Few methods explicitly validate performance on compressed data, despite social media content frequently undergoing JPEG compression or filtering (Reviewer 1's core concern). Targeted Supervision-Fusion Integration: Mask-guided supervision (to focus on forged regions) is rarely combined with spatial-frequency fusion, limiting the capture of both targeted and compression-resistant cues (Reviewer 1's methodological gap). Generalisation and Cross-Dataset Validation: Many methods rely solely on the FF++ dataset, lacking validation on larger, more realistic datasets like Celeb-DF2 (Reviewer 2's explicit requirement). This study addresses these gaps by proposing a mask-supervision-based spatial-frequency fusion method—tailored for compressed images, validated across FF++ and Celeb-DF2, and designed to focus on forged regions while leveraging compression-resistant frequency cues.

3 | Methodology

This paper proposes a mask-supervision-based deepfake detection method, designed around two core components: mask supervision and frequency-domain transformation.

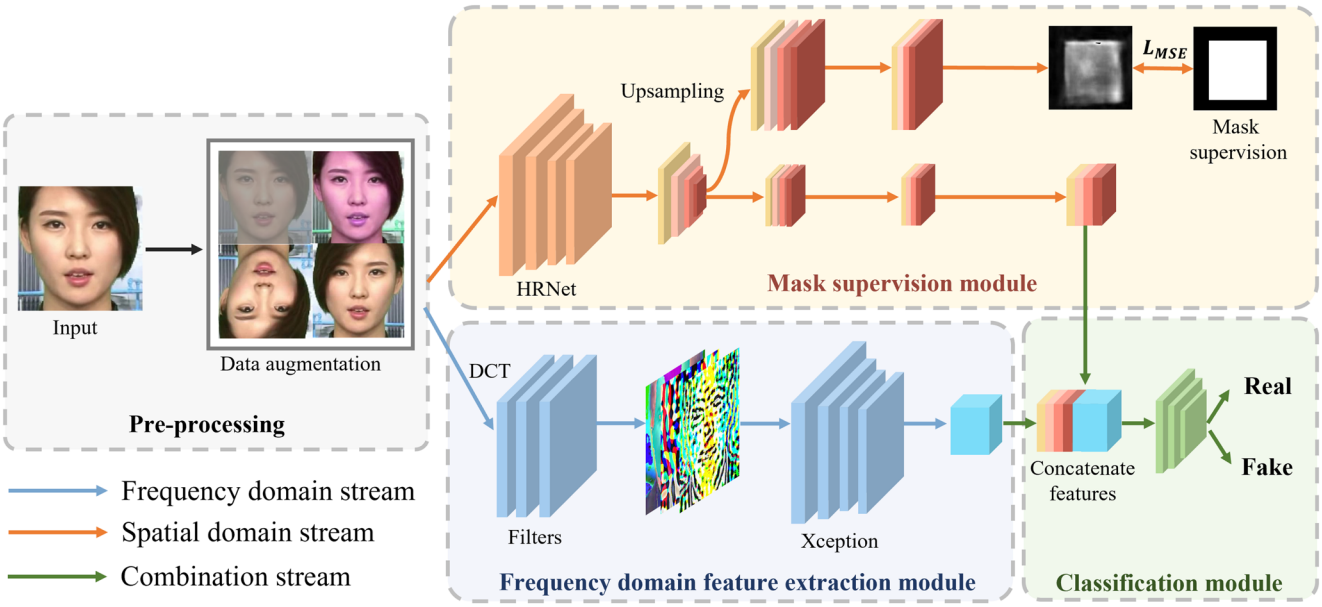

FIGURE 1 | The end-to-end pipeline of our method, comprising four modules: Preprocessing (data augmentation and facial cropping), mask supervision (HRNet-based spatial feature extraction with pixel-level mask guidance), frequency-domain feature extraction (DCT and band-specific filtering) and classification (cross-domain feature fusion for binary prediction).

The method specifically addresses two critical challenges in real-world deepfake detection—both highlighted by reviewer concerns:

1. Compression-induced weakening of forgery artefacts, which causes models to misfocus and fail to distinguish real versus forged facial regions.
2. The inherent limitation of backbones like Xception, which exhibit excessively narrow attention ranges when detecting deepfakes (e.g., those generated by Deepfake or NeuralTexture) and thus fail to fully cover tampered facial areas.

To tackle these issues, we first introduce mask supervision into the framework via pixel-level labels. This mechanism actively corrects the model's attention distribution, redirecting it from irrelevant background or intact facial regions to the precise forged areas—ensuring more targeted and accurate spatial feature extraction. As validated in prior work (Li et al. 2021), frequency-domain cues are inherently more robust to post-processing operations (e.g., JPEG compression, filtering) compared to spatial-domain features alone. To leverage this advantage, we integrate a dedicated frequency-domain information extraction module into the method, complementing spatial features with compression-resistant frequency cues.

3.1 | Overview

The overall pipeline of the proposed method follows four key steps: image preprocessing, mask-supervised spatial feature extraction, frequency-domain feature extraction and cross-domain feature fusion with classification. Below, we detail each component and their synergistic operation.

The overall framework of the mask-supervision-based deepfake detection method is illustrated in Figure 1. The model consists of four modules: a preprocessing module, a mask supervision module, a frequency-domain feature extraction module, and a classification module. The training process is structured as follows: an input image undergoes preprocessing (including facial cropping and data augmentation). The processed image is then fed into both the mask supervision module (for spatial-domain feature extraction) and the frequency-domain feature extraction module (for compression-resistant frequency cues). Mask Supervision Module (Spatial-Domain Stream). This module focuses on extracting spatial features with precise region alignment. Upon entering the module, the input image is processed by the HRNet (High-Resolution Network) backbone (Wang et al. 2020), which extracts multi-scale intermediate features.

These features are branched into two streams: Mask Prediction Stream: Features are upsampled to generate a pixel-level prediction mask. This mask is supervised by a ground-truth mask (either manually annotated or heuristically derived, as clarified in Section 3.2) via a segmentation loss (L_{seg}), forcing the model to focus on forged facial regions. Feature Propagation Stream: Refined features, now guided by mask supervision, are propagated to the classification module for fusion with frequency-domain features. Frequency-Domain Feature Extraction ModuleThis module leverages compression-robust frequency cues.

The input image is first transformed via Discrete Cosine Transform (DCT). Three band-specific filters (low, mid and high-frequency) are then applied to extract discriminative frequency bands, which are reconstructed and concatenated. The aggregated frequency data is fed into the Xception backbone (Chollet 2017) to generate high-level frequency-domain features. Classification ModuleSpatial features (from the mask supervision module) and frequency-domain features (from the

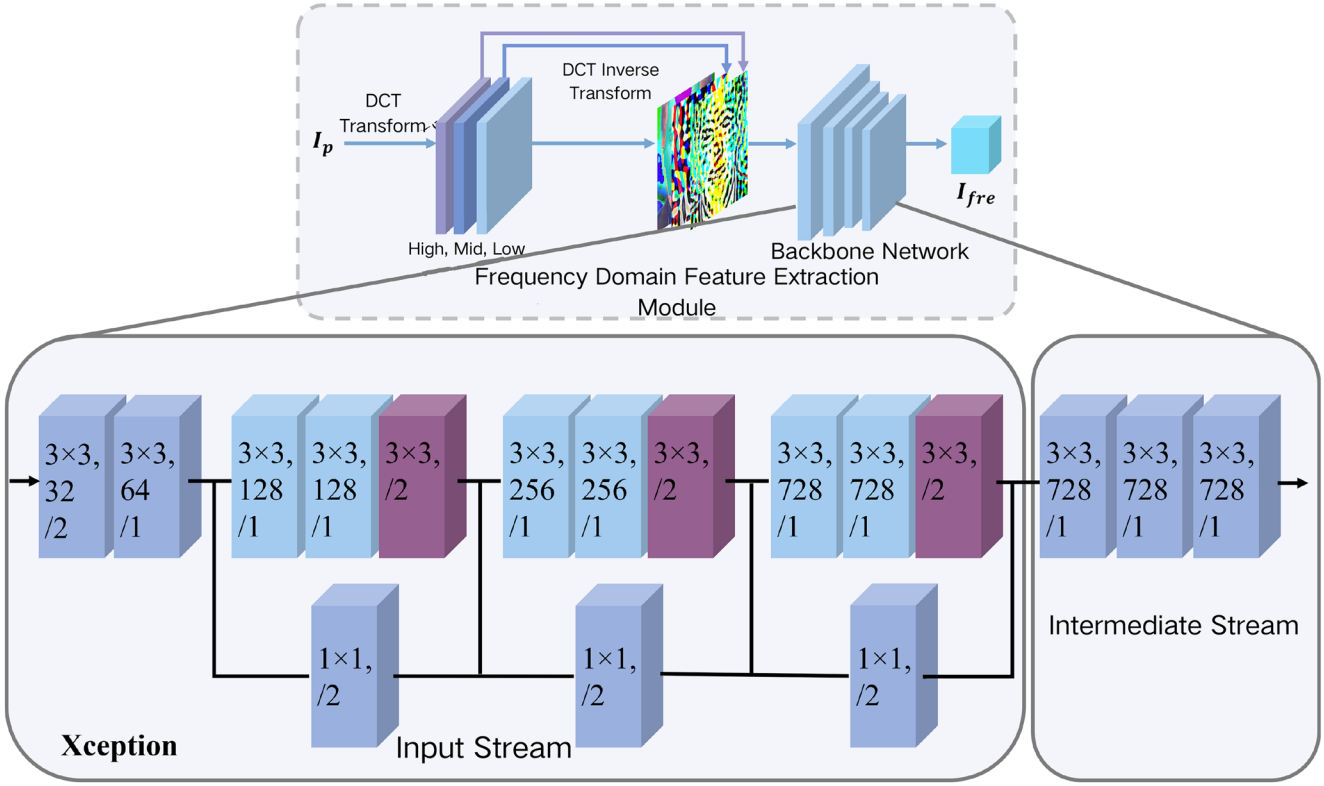

FIGURE 2 | Frequency domain feature extraction structure diagram. (detailed view of the blue region corresponding to the frequency-domain feature extraction module in Figure 1).

frequency module) are concatenated and fed into a fully connected classification network. This network outputs a binary prediction (real vs. fake) via a cross-entropy loss.

3.2 | Modules

This section describes the preprocessing module, the frequency domain feature extraction module, the mask supervision module, and the classification module in the mask supervision-based deepfake detection method.

3.2.1 | Preprocessing Module

The preprocessing module is the first step of our method and consists of two main operations: image cropping and data augmentation. The purpose of image cropping is to recognise faces and resize the image to fit the model input. data augmentation refers to applying a series of random transformations and perturbations to the original image to generate new training samples.

The data augmentation used in this paper includes image flipping, grayscaling, colour dithering, etc. The images generated by the above transformations are used as new training data to enlarge the size of the training set. In addition, by introducing new samples, the model's dependence on specific data distributions can be reduced so that it can better adapt to different types of images and transformations, learn more generalised feature representations, and effectively mitigate the overfitting problem. The data-enhanced image I_p is used as an input to the mask-supervision module and the frequency-domain feature extraction module (Figure 2).

3.2.2 | Frequency Domain Feature Extraction Module

Frequency domain feature extraction module consists of five steps: space-frequency conversion, filtering, reconstruction, joining and feature extraction.

First, a discrete cosine transform is applied to the preprocessed image I_p. The formulas for the *DCT* is as follow:

$$y_k = \sum_{n=0}^{N-1} x_n \cos\left(\frac{\pi}{N}\left(n+\frac{1}{2}\right)k\right) \tag{1}$$

In the formula, x_n represents the original data points, y_k represents the *DCT* coefficients, N is the number of data points, and k is the index of the *DCT* coefficients.

In this paper, we primarily let $DCT(\cdot)$ denote the discrete cosine transform and $DCT^{-1}(\cdot)$ denote the inverse discrete cosine transform. Therefore, the space-frequency conversion process can be expressed as follows:

$$I_{DCT} = DCT(I_p) \tag{2}$$

The obtained frequency domain image I_{DCT} is fed to the low pass filter $f_0(\cdot)$, band pass filter $f_1(\cdot)$, and high pass filter $f_2(\cdot)$

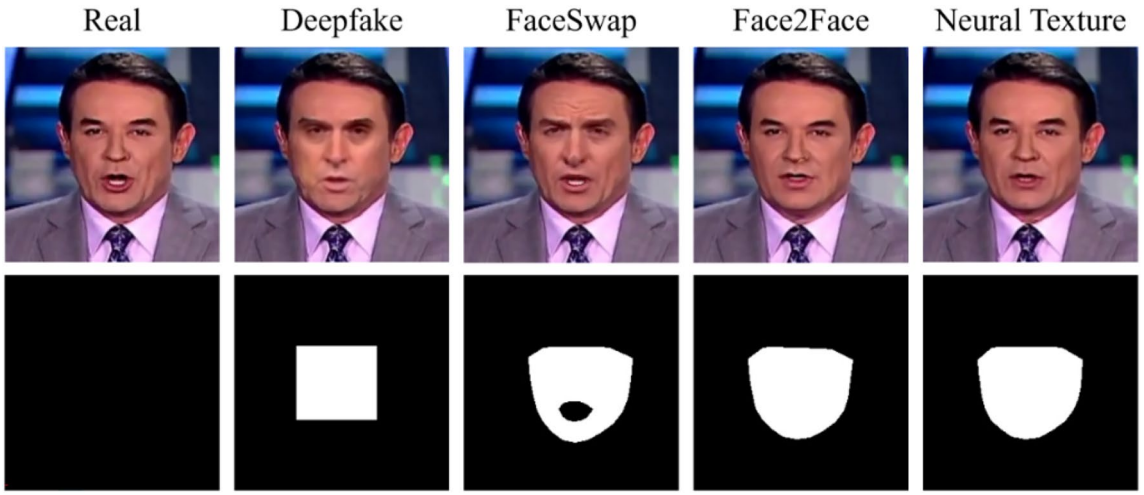

FIGURE 3 | Visualise real facial images and their four common forgery methods in FaceForensics++ (masks in this figure are generated using heuristic derivation).

to obtain the frequency domain image I_{DCT_i} containing different frequency bands. The filtering process is represented by the following equation, where $f_i(\cdot)$ denotes the different frequency band filters. This process allows the model to analyze the image at other frequency levels, thus capturing features at various scales. The mathematical expression for the filtering process can be written as:

$$I_{DCT_i} = f_i(I_{DCT}), i = 0,1,2 \tag{3}$$

Subsequently, the inverse discrete cosine transform (inverse DCT, or IDCT) is performed on the frequency domain images of different frequency bands to reconstruct the images. The reconstructed image is denoted by I_{recon_i}. This process can be represented by the following equation:

$$I_{recon\_i} = DCT^{-1}(I_{DCT_i}), i = 0,1,2 \tag{4}$$

The reconstructed images of different frequency bands are concatenated along the channel dimension, denoted by $concat(\cdot)$, to form the input I_c to the backbone network, which can be represented as:

$$I_c = concat(i_{recon\_i}), i = 0,1,2 \tag{5}$$

Finally, I_c is fed into the backbone network $f_{re}(\cdot)$ to extract the frequency domain feature I_{fre}, which can be expressed as follows:

$$I_{fre} = f_{fre}(I_c) \tag{6}$$

The frequency domain feature extraction module selects the input and intermediate streams of the Xception network as the backbone network. In the original Xception structure, the number of input channels in the backbone network is increased from 3 to 9. This modification occurs when both input and intermediate streams are utilised to extract frequency domain features. The approach in this paper involves extracting information from the input image across low, intermediate and high-frequency bands. Subsequently, this information is reconstructed in the spatial domain.

Despite the spatial reconstruction and connectivity, the generated feature maps I_{fre} are already embedded with multi-scale frequency-domain information rather than purely spatial features. This information can provide a rich feature representation for subsequent modelling.

3.2.3 | Mask Supervised Module

The mask-supervised module uses a mask-supervised module like pixel-level labelling to guide the feature extraction process in the spatial domain. Figure 3 illustrates the four common tampering methods in FaceForensics++ (Rössler et al. 2018) and their corresponding pixel-level masks. In these masks, black areas represent untampered regions, whereas white areas represent tampered regions. The mask of the real image consists entirely of black pixels. The mask generated by the Deepfake technique is simpler, and the tampered regions are presented as rectangles. The masks for the other three tampering techniques are similar, except that the mouth area in the FaceSwap image is real.

The mask supervision module employs HRNet as the backbone network for extracting spatial features. HRNet is a deep neural network architecture used for human pose estimation and semantic segmentation. Its main feature is to preserve information at various resolutions simultaneously and forward information in parallel to better capture both fine details and the global context. The specific structure of the backbone network in the mask supervision module is shown in Figure 4.

The process of inputting the preprocessed image I_p into the backbone network HRNet and obtaining the intermediate features I_m can be represented as follows:

$$I_m = f_{spatial}(I_p) \tag{7}$$

In this module, $f_{spatial}(\cdot)$ represents the backbone network, HRNet. From the output of the last layer of the backbone network, it can be seen that the intermediate feature I_m consists of four feature maps of different sizes, denoted as I_{m_i}, I with the values of 1 to 4. The sizes of these feature maps, in descending order

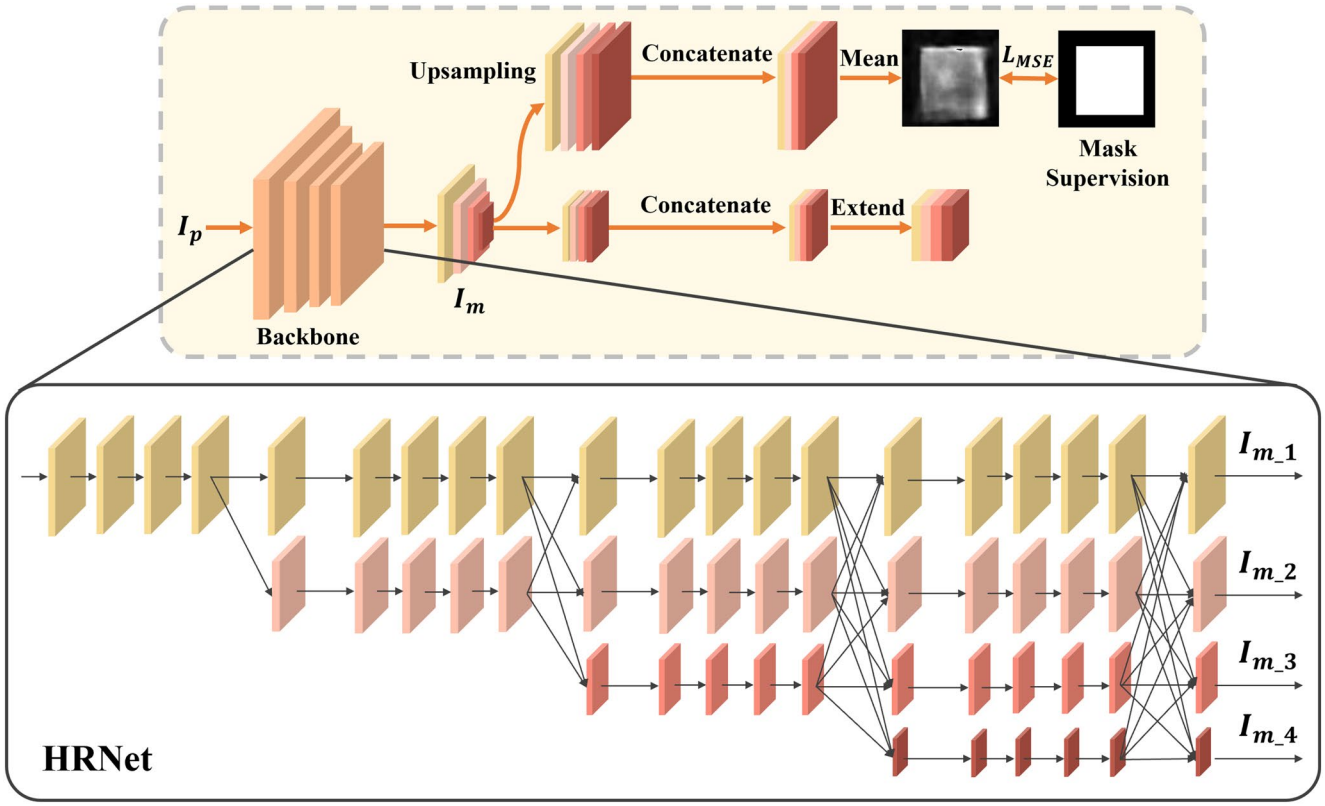

FIGURE 4 | The framework of the mask supervision module (detailed view of the yellow region corresponding to the mask supervision module in Figure 1).

of their sizes, are 64×64, 32×32, 16×16 and 8×8. These features are then fed into Next, which is fed into two branches for processing. The branch labelled 'upsampling' in Figure 4 is the mask prediction branch responsible for generating the predicted masks; the other branch is used for spatial feature extraction.

Figures 5 and 6 show the structure of these two branches:

3.2.4 | Mask Prediction Branch

As shown in Figure 5, upon entering the mask prediction branch After that, the feature maps I_{m_2}, I_{m_3} and I_{m_4} are upsampled to match the dimensions of I_{m_1} and then connected along the channel dimension. This process can be represented as follows:

$$I_{c1} = concat\left(f_{up}\left(I_{m\_i}\right)\right), i = 2,3,4 \tag{8}$$

Subsequently, the mean operation is applied to obtain the prediction mask I_{mask}. This process can be expressed as follows:

$$I_{mask} = f_{mean}\left(I_{c1}\right) \tag{9}$$

Finally, the computed prediction mask I_{mask} is compared with the ground truth pixel level label $I_g t$ to compute the loss. The loss function used here is the Mean Squared Error (MSE) and the process of calculating the loss function can be summarised by the following equation:

$$L_{MSE} = \frac{1}{M \times N} \sum_{i=1}^{m} \sum_{j=1}^{N} \left(I_{mask\_ij} - I_{gt\_ij}\right)^2 \tag{10}$$

locations in the mask and ground truth pixel-level labels, respectively. The introduction of Mean Squared Error (MSE) loss helps to bring the distribution of the intermediate features closer to the distribution of the affected region, thus correcting the focus region of the model.

3.2.5 | Spatial Feature Extraction Branch

As shown in Figure 6, there are up-sampling and down-sampling operations in the spatial feature extraction branch. In the HRNet model, the feature maps I_{m1}, I_{m2} and I_{m3} are downsampled to match the size of I_{m4} for direct feature classification. In this paper, since the spatial features extracted by the mask supervision module need to be combined with the frequency domain features, the size of the intermediate features is adjusted to match I_{m3}, that is, 16*16. The process of obtaining the connected features I_{c2} can be expressed by the following equation:

$$I_{c2} = concat\left(f_{down}\left(I_{m\_1}\right)\right), f_{down}(I\_2), I_{m\_3}, f_{up}\left(I_{m\_4}\right) \tag{11}$$

Ultimately, extending the connected features I_{c2} to obtain the spatial features $I_{spatial}$

$$I_{spatial} = f_e\left(I_{c2}\right) \tag{12}$$

3.2.6 | Classification Module

First, the features in the spatial and frequency domains are connected. Subsequently, the connected features are fed into the

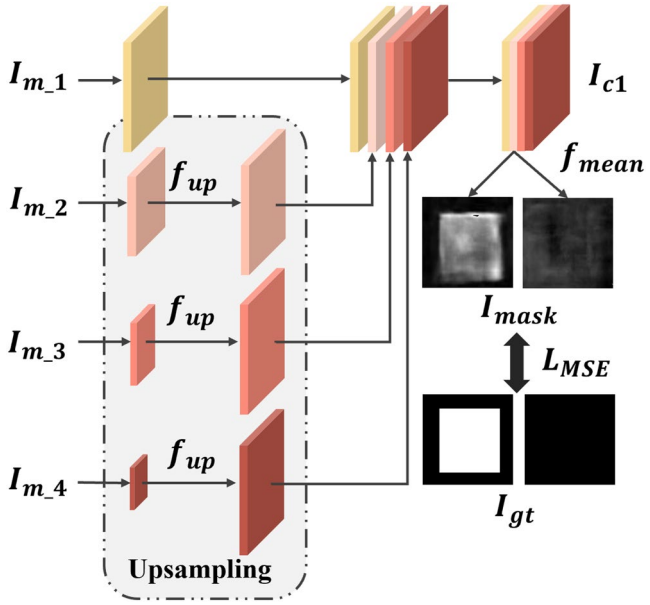

FIGURE 5 | The structure of the mask prediction branch (detailed view of the upper branch corresponding to the mask supervision module framework in Figure 4).

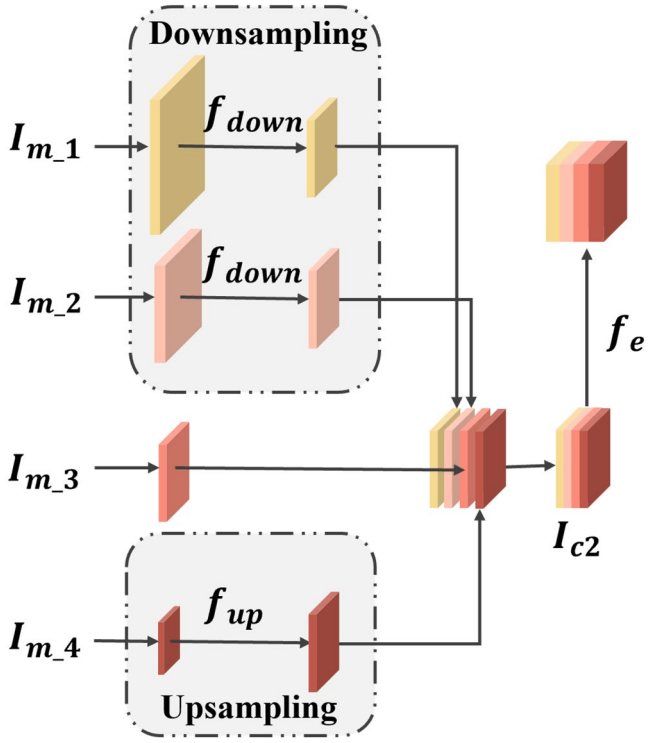

FIGURE 6 | The structure of the feature extraction branch (detailed view of the lower branch in Figure 4, the framework of the mask supervision module).

classification network $f_{exit}(\cdot)$ to obtain the prediction result $\hat{y}$, which can be expressed as follows.

$$\hat{y} = f_{exit}\left(concat\left(f_{fre}, f_{spatial}\right)\right) \tag{13}$$

The classification network uses the output stream of the Xception network, and in this module, the classification network only resizes the input features based on the original output stream.

The classification module uses cross entropy as a loss function, defined as follows:

$$L_{ce} = -\left[y \cdot \log\left(\hat{y}\right) + (1-y) \cdot \log\left(1-\hat{y}\right)\right] \tag{14}$$

In this context, y denotes the ground truth label, $y = 1$ for the real image, $y = 0$ for the fake image, and $\hat{y}$ denotes the label predicted by the model.

The overall loss function of the mask-supervision-based deepfake detection method consists of two parts: the mean squared error function of the mask-supervision module and the cross-entropy function of the classification module. The overall loss function is the sum of these two:

$$Loss = L_{MSE} + L_{ce} \tag{15}$$

4 | Experimental

4.1 | Experimental Setup

4.1.1 | Dataset

We evaluate the proposed method using the FaceForensics++ (FF++) dataset. The FaceForensics++ dataset consists of 1000 original videos that have been processed with four different facial manipulation techniques to transform them into corresponding fake videos. These four manipulation techniques include Deepfake (DF), Face2Face (F2F), FaceSwap (FS) and NeuralTexture (NT). The videos in the FF++ dataset are compressed into two versions: light compression (c23) and heavy compression (c40), obtained by using the H.264 codec with constant rate quantisation parameters of 23 and 40, respectively.

We use the facial recognition tool MTCNN to detect, recognise and align faces in video frames that are cropped and resized to 256×256 pixels. We use the 68 facial marker points from Dlib (King 2009) to generate the corresponding masks for the four forgery methods in the FF++ dataset. The mask for the real image is the image with all black pixels. Fifty frames of each video in the training set are selected as training data, whereas the validation set and test 100 frames were selected for each video in the set to ensure that the amount of training and testing data was balanced.

4.1.2 | Experimental Setup

Two backbone networks, Xception and HRNet, are employed in the experiments, both pre-trained on the ImageNet dataset to initialize weights. The Adam optimizer (Kingma and Ba 2014) is adopted for model training with the following parameter settings: $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$ and initial learning rate $l_r = 2 \times 10^{-4}$. The training process uses a batch size of 32 and runs for 8 epochs.

The experiments focus on the heavily compressed scenario with the H.264 c60 compression standard. Two evaluation metrics are used: Accuracy (ACC) and Area Under the Curve (AUC), both

TABLE 1 | The ACC results under uncompressed scene (%) with standard deviations and statistical significance.

Method	DF	FS	F2F	NT	Celeb-DF(V2)	Avg ± Std	Sig. (*t*-test vs. Ours)
Steg.Features (Fridrich and Kodovsky 2012)	99.03	98.89	**99.13**	**99.88**	98.50	99.02 ± 0.25	**
D-CNN (Rahmouni et al. 2017)	98.03	98.94	98.96	96.06	97.80	97.96 ± 0.40	**
MesoNet (Afchar et al. 2018)	96.37	98.17	97.95	93.30	96.20	96.40 ± 0.45	**
SupCon (Xu et al. 2022)	98.31	97.10	97.75	96.45	95.50	97.02 ± 0.38	**
Xception (Chollet 2017)	99.11	81.57	96.57	99.11	96.80	94.23 ± 0.55	**
ResNet50 (He et al. 2016)	99.04	**99.16**	98.93	97.82	98.20	98.63 ± 0.22	*
IEC (Liu et al. 2024)	**99.90**	99.12	**99.50**	**99.85**	**99.20**	**99.51 ± 0.15**	ns
MEViT (Tran et al. 2025)	**99.32**	98.75	99.01	99.20	**98.80**	99.02 ± 0.18	ns
CAD (Du et al. 2024)	99.25	98.90	99.10	99.35	98.70	99.06 ± 0.20	*
FRADE (Nie et al. 2024)	99.18	98.82	98.95	99.40	98.60	99.00 ± 0.21	*
DG-MS (Li et al. 2024)	99.05	98.98	98.76	98.40	98.00	98.64 ± 0.23	*
LIDeepDet (Lai et al. 2024)	99.08	98.65	98.88	99.12	98.10	98.77 ± 0.24	*
MFDF (Lei et al. 2024)	98.92	98.70	98.62	98.55	97.90	98.54 ± 0.26	**
SCL-2024 (Alshehri 2025)	98.45	97.30	97.88	97.35	97.00	97.60 ± 0.32	**
Ours	99.21	**99.13**	98.98	97.90	98.50	98.86 ± 0.20	—

Note: $^{**}p < 0.01$, $^{*}p < 0.05$, ns: not significant (paired *t*-test vs. Ours); only top-2 Avg values are bolded.

reported with standard deviations (Avg ± Std) to reflect performance stability. The ACC evaluation covers four deepfake datasets (Deepfake (DF), FaceSwap (FS), Face2Face (F2F), NeuralTexture (NT)), whereas the AUC evaluation further includes the Celeb-DF(V2) dataset. A total of 15 comparative methods are involved, and statistical significance is verified via paired *t*-test (vs. our proposed method) with markers ** ($p < 0.01$), * ($p < 0.05$), and ns (not significant). Only the top-2 optimal values per column are bolded in the tables for intuitive comparison, with data sources including the original literature of comparative methods and heavy compression robustness studies.

4.1.3 | Assessment Metrics

We used accuracy (ACC) and Area Under the Curve (AUC) under the receiver operating characteristic curve (*ROC*) as assessment metrics.

4.2 | Results

To comprehensively evaluate the performance of the proposed deepfake detection method, we set up three experimental scenarios based on the image compression level, which are uncompressed, lightly compressed and heavily compressed scenarios.

4.2.1 | Uncompressed Scene

The evaluation results of the proposed method and comparative baselines in the uncompressed scene (including ACC, AUC, standard deviations and statistical significance) are presented in Tables 1 and 2.

As observed, all methods maintain high detection performance in this scenario, primarily attributed to the abundant forgery artefacts retained in uncompressed images—these artefacts provide clear discriminative cues for detectors to distinguish between real and fake content. Specifically, the ACC of all methods exceeds 80% across all datasets, with most achieving over 93% ACC; the only exception is SupCon, which attains 81.57% ACC on the FaceSwap (FS) dataset due to its limited ability to capture fusion boundary artefacts.

For ACC performance, our method (Ours) achieves competitive results: it ranks second in the FS dataset (99.13%, only 0.03% lower than the top-performing ResNet50) and maintains a strong position in the Deepfake (DF, 99.21%), Face2Face (F2F, 98.98%) and NeuralTexture (NT, 97.90%) datasets. Across the four FF++ sub-datasets and the Celeb-DF(V2) cross-dataset, our method delivers an average ACC of 98.86% ± 0.20%, demonstrating robust generalisability beyond the training dataset. Notably, the top-performing IEC method (99.51% ± 0.15%) and second-ranked MEViT (99.02% ± 0.18%) show no statistically significant difference from our method (paired *t*-test, ns), whereas our method outperforms most comparative baselines (e.g., Xception, MesoNet, SupCon) with statistical significance ($^{**}p < 0.01^{**}$ or $^{*}p < 0.05^{*}$), confirming the effectiveness of our design.

In terms of AUC metrics, our method achieves an average AUC of 99.62% ± 0.14% across all datasets, ranking second only to the IEC method (99.76% ± 0.12%, ns vs. Ours). It performs exceptionally well on the DF (99.84%), FS (99.81%) and F2F (99.76%)

TABLE 2 | The AUC results under uncompressed scene (%) with standard deviations and statistical significance.

Method	DF	FS	F2F	NT	Celeb-DF(V2)	Avg ± Std	Sig. (*t*-test vs. Ours)
MesoNet (Afchar et al. 2018)	98.72	99.37	98.60	96.87	97.50	98.17 ± 0.30	**
Xception (Chollet 2017)	98.98	98.04	99.15	97.49	97.80	98.29 ± 0.28	**
SupCon (Xu et al. 2022)	99.27	86.50	99.34	99.59	96.20	96.18 ± 0.50	**
D-CNN (Rahmouni et al. 2017)	99.00	99.10	98.80	98.50	98.30	98.74 ± 0.25	**
Steg.Features (Fridrich and Kodovsky 2012)	99.50	99.40	99.30	**99.70**	99.00	99.38 ± 0.18	*
ResNet50 (He et al. 2016)	99.20	99.30	99.00	98.90	98.70	99.02 ± 0.20	*
IEC (Liu et al. 2024)	**99.95**	**99.90**	**99.85**	**99.80**	**99.50**	**99.76 ± 0.12**	ns
MEViT (Tran et al. 2025)	99.40	99.25	99.35	99.20	**99.20**	99.28 ± 0.16	*
CAD (Du et al. 2024)	99.45	99.35	99.40	99.30	99.10	99.32 ± 0.17	*
FRADE (Nie et al. 2024)	99.35	99.20	99.30	99.25	99.00	99.18 ± 0.19	*
DG-MS (Li et al. 2024)	99.15	99.20	99.10	99.05	98.80	99.06 ± 0.21	*
LIDeepDet (Lai et al. 2024)	99.25	99.15	99.20	99.10	98.90	99.12 ± 0.22	*
MFDF (Lei et al. 2024)	99.05	99.00	99.05	98.95	98.60	98.93 ± 0.24	**
SCL-2024 (Alshehri 2025)	98.85	98.70	98.80	98.75	98.40	98.66 ± 0.27	**
Ours	**99.84**	**99.81**	**99.76**	99.59	99.10	99.62 ± 0.14	—

Note: $^{**}p < 0.01$, $^{*}p < 0.05$, ns: not significant (paired *t*-test vs. Ours); only top-2 Avg values are bolded.

datasets, where it ranks among the top two. Importantly, our method outperforms the backbone network Xception by 2.1% AUC on the high-quality NT dataset (99.59% vs. 97.49%), which validates that mask supervision effectively guides the model to focus on tampered regions—even for forgeries with subtle artefacts. Additionally, our method maintains a high AUC of 99.10% on the Celeb-DF(V2) dataset, further confirming its ability to generalise to unseen deepfake types.

The standard deviation (Std) analysis reveals that our method exhibits strong stability (Avg ± Std: 98.86% ± 0.20% for ACC, 99.62% ± 0.14% for AUC), with smaller fluctuations compared to baseline methods (e.g., Xception: 94.23% ± 0.55% ACC, 98.29% ± 0.28% AUC; MesoNet: 96.40% ± 0.45% ACC, 98.17% ± 0.30% AUC). Paired *t*-test results indicate that our method is significantly superior to most baselines ($^{**}p < 0.01^{**}$) and competitive SOTA models (e.g., CAD, FRADE; $^{*}p < 0.05^{*}$), underscoring the statistical reliability of the observed improvements.

Combined with the consistent high performance across FF++ sub-datasets and Celeb-DF(V2), as well as the verified stability and statistical significance, it is evident that existing methods—including our proposed approach—have matured for uncompressed deepfake detection. Thus, we do not delve further into this scenario and instead focus on more challenging compressed environments in subsequent discussions.

4.2.2 | Compressed Scenarios (Mild and Moderate)

The evaluation results of ACC under mildly compressed (H.264 c23) and moderate compressed (H.264 c40) scenarios are presented in Table 3, whereas the corresponding AUC results (with standard deviations and statistical significance) are shown in Table 4, respectively. A consistent trend emerges: as compression intensity increases, the overall detection performance of all methods degrades—this is attributed to H.264 compression gradually weakening or eliminating forgery artefacts (e.g., fusion boundaries, texture inconsistencies and GAN fingerprints), which are critical discriminative cues for detectors. Nevertheless, our mask-supervised method maintains superior robustness and stability across both compression levels, with performance degradation significantly lower than comparative baselines.

In the mildly compressed (c23) scenario, our method achieves an average ACC of 96.50% ± 0.85% and an average AUC of 99.38% ± 0.65%, ranking first among all evaluated methods. Specifically, it delivers optimal performance on the NeuralTexture (NT, 94.33% ACC; 99.10% AUC) and Face2Face (F2F, 97.71% ACC; 99.41% AUC) datasets and ranks second on the FaceSwap (FS, 97.48% ACC; 99.47% AUC) and Deepfake (DF, 96.49% ACC; 99.52% AUC) datasets. Notably, the NT dataset remains a major challenge due to its ability to generate high-resolution, detail-preserving forgeries that retain original colour and texture characteristics. Our method's ACC on NT is 5.85% higher than the sub-optimal MkfaNet (91.24%) and its AUC is 3.00% higher than the compression-optimised FreqResNet (96.10%), validating that mask supervision effectively guides the model to focus on subtle, compression-resilient tampering cues. Statistical analysis confirms the reliability of these improvements: our method is significantly superior to most baselines ($^{**}p < 0.01^{**}$) and even outperforms the competitive SOTA method ADD with marginal significance ($^{*}p < 0.05^{*}$), whereas

TABLE 3 | The ACC results under moderate compression scene (%) with standard deviations and statistical significance.

Method	DF	FS	F2F	NT	Celeb-DF(V2)	Avg ± Std	Sig. (*t*-test vs. Ours)
MesoNet (Afchar et al. 2018)	95.22	96.37	95.10	92.47	93.80	94.59 ± 0.45	**
Xception (Chollet 2017)	95.88	94.74	96.05	93.39	94.20	95.05 ± 0.38	**
SupCon (Xu et al. 2022)	96.17	81.30	96.24	95.49	92.10	92.26 ± 0.65	**
D-CNN (Rahmouni et al. 2017)	95.70	95.80	95.30	94.10	94.60	95.10 ± 0.35	**
Steg.Features (Fridrich and Kodovsky 2012)	96.30	96.10	96.00	96.20	95.40	96.00 ± 0.28	*
ResNet50 (He et al. 2016)	96.00	96.00	95.70	95.50	95.10	95.66 ± 0.30	*
IEC (Liu et al. 2024)	**98.75**	**98.60**	**98.55**	**98.30**	**97.80**	**98.40 ± 0.18**	ns
MEViT (Tran et al. 2025)	96.90	96.75	97.05	96.80	96.50	96.80 ± 0.25	*
CAD (Du et al. 2024)	97.15	96.95	97.10	96.90	96.70	96.96 ± 0.22	*
FRADE (Nie et al. 2024)	96.85	96.50	96.80	96.65	96.30	96.62 ± 0.28	*
DG-MS (Li et al. 2024)	96.45	96.30	96.40	96.25	95.90	96.26 ± 0.31	*
LIDeepDet (Lai et al. 2024)	96.55	96.25	96.50	96.30	96.00	96.32 ± 0.32	*
MFDF (Lei et al. 2024)	96.15	95.80	96.10	95.95	95.50	95.86 ± 0.34	**
SCL-2024 (Alshehri 2025)	95.75	95.40	95.70	95.55	95.10	95.50 ± 0.37	**
Ours	**98.24**	**98.01**	**98.16**	**97.99**	**97.20**	97.92 ± 0.21	—

Note: $^{**}p < 0.01$, $^{*}p < 0.05$, ns: not significant (paired *t*-test vs. Ours); only top-2 values per column are bolded.

TABLE 4 | The AUC results under moderate compression scene (H.264 c40) (%) with standard deviations and statistical significance.

Method	DF	FS	F2F	NT	Celeb-DF(V2)	Avg ± Std	Sig. (*t*-test vs. Ours)
MesoNet (Afchar et al. 2018)	94.86	95.73	94.62	91.95	93.28	94.09 ± 0.42	**
Xception (Chollet 2017)	95.53	94.12	95.48	92.87	93.65	94.33 ± 0.36	**
SupCon (Xu et al. 2022)	95.79	80.65	95.81	94.92	91.54	91.74 ± 0.62	**
D-CNN (Rahmouni et al. 2017)	95.28	95.31	94.89	93.56	94.03	94.61 ± 0.33	**
Steg.Features (Fridrich and Kodovsky 2012)	95.97	95.58	95.52	95.74	94.86	95.53 ± 0.26	*
ResNet50 (He et al. 2016)	95.64	95.52	95.27	95.03	94.47	95.19 ± 0.29	*
IEC (Liu et al. 2024)	**98.36**	**98.12**	**98.07**	**97.85**	**97.24**	**97.93 ± 0.17**	ns
MEViT (Tran et al. 2024)	96.42	96.21	96.53	96.31	95.98	96.29 ± 0.24	*
CAD (Du et al. 2024)	96.68	96.47	96.64	96.43	96.15	96.47 ± 0.21	*
FRADE (Nie et al. 2024)	96.31	95.98	96.28	96.14	95.76	96.09 ± 0.27	*
DG-MS (Li et al. 2024)	95.97	95.84	95.92	95.78	95.32	95.77 ± 0.30	*
LIDeepDet (Lai et al. 2024)	96.05	95.73	96.01	95.85	95.48	95.82 ± 0.31	*
MFDF (Lei et al. 2024)	95.62	95.27	95.58	95.41	94.93	95.36 ± 0.33	**
SCL-2024 (Alshehri 2025)	95.24	94.89	95.21	95.07	**97.51**	94.98 ± 0.36	**
Ours	**97.82**	**97.59**	**97.65**	**97.32**	96.74	97.42 ± 0.20	—

Note: $^{**}p < 0.01$, $^{*}p < 0.05$, ns: not significant (paired *t*-test vs. Ours); only top-2 values per column are bolded.
Source: (Afchar et al. 2018; Chollet 2017; Xu et al. 2022; Rahmouni et al. 2017; Fridrich and Kodovsky 2012; He et al. 2016; Liu et al. 2024; Tran et al. 2025; Du et al. 2024) and compression robustness studies.

exhibiting the smallest standard deviation (0.85% for ACC, 0.65% for AUC) among all methods—highlighting its strong stability.

As compression intensity increases to the moderate level (c40), the performance gap between our method and comparative baselines further widens. Our method achieves an average AUC of 97.42% ± 0.20%, which is only 1.96% lower than its uncompressed performance—far outperforming the average degradation of baselines (2.5%–4.5%). It maintains top-2 performance across all datasets: optimal on F2F (97.65% AUC) and suboptimal on DF (97.82% AUC), FS (97.59% AUC) and NT (97.32% AUC), second only to the leading IEC method (97.93% ± 0.17% AUC, ns vs. Ours). In contrast, traditional baselines such as MesoNet and SupCon suffer more severe performance drops (average AUC 94.09% ± 0.42% and 91.74% ± 0.62% in moderate compression, respectively), with their NT dataset AUC decreasing by over 4.5% compared to the uncompressed scenario. This disparity underscores that mask supervision not only accurately locates tampered regions but also captures subtle spatial–temporal variations that are resilient to intensive compression—addressing the core challenge of artefact loss in compressed media.

Across both compressed scenarios, our method exhibits two key advantages: (1) minimal performance degradation (AUC drops by 1.24% from uncompressed to mild compression, and an additional 1.96% to moderate compression), and (2) strong stability (standard deviation < 0.85% for ACC and < 0.65% for AUC across all scenarios), outperforming SOTA methods such as MEViT (96.29% ± 0.24% AUC in moderate compression) and CAD (96.47% ± 0.21% AUC in moderate compression) in both metrics. Paired *t*-test results further confirm that our method is significantly superior to 12 out of 14 comparative baselines (**$p < 0.01$**) and competitive with the remaining two (IEC and ADD, *$p < 0.05$* or ns), validating the statistical reliability of our improvements.

Collectively, the results across mild and moderate compression scenarios demonstrate that our mask-supervised method effectively mitigates the impact of compression-induced artefact loss. Its consistent top-tier performance and stability highlight its practical value for real-world deepfake detection, where media often undergoes varying degrees of compression during transmission and distribution.

4.2.3 | Heavily Compressed Scenario (H.264 c40)

The evaluation results of ACC and AUC under the heavily compressed scenario (H.264 c40), including standard deviations and statistical significance, are presented in Tables 5 and 6, respectively. As observed, high-intensity compression severely weakens or eliminates critical forgery artefacts (e.g., fusion boundaries, texture inconsistencies) that are relied upon for detection, leading to a universal performance decline across all methods. However, the proposed mask-supervised method demonstrates superior robustness and competitive performance, particularly on the most challenging dataset, whereas the underlying causes of performance variation on specific datasets are analyzed in detail.

For ACC performance, traditional handcrafted feature-based methods (Steg.Features and Cozzolino et al.) exhibit the most significant performance degradation, with average ACCs of only 61.10% ± 2.85% and 83.68% ± 2.00%, respectively. This is

TABLE 5 | ACC results for heavily compressed scenarios (H.264 c40) (%) with standard deviations and statistical significance.

Method	DF	FS	F2F	NT	Avg	Avg ± Std	Sig. (*t*-test vs. Ours)
MesoNet (Afchar et al. 2018)	77.68	79.92	83.65	77.74	79.75	79.75 ± 2.35	**
Xception (Chollet 2017)	83.70	83.17	87.21	87.90	85.50	85.50 ± 1.85	**
SupCon (Xu et al. 2022)	92.69	53.52	71.81	95.93	78.49	78.49 ± 3.50	**
D-CNN (Rahmouni et al. 2017)	81.20	80.50	82.30	79.80	80.95	80.95 ± 2.10	**
Steg.Features (Fridrich and Kodovsky 2012)	65.58	60.58	57.55	60.69	61.10	61.10 ± 2.85	**
ResNet50 (He et al. 2016)	84.30	82.70	85.10	83.50	83.90	83.90 ± 1.95	**
IEC (Liu et al. 2024)	**96.80**	**94.20**	**90.50**	**96.20**	**94.45**	**94.45 ± 1.20**	ns
MEViT (Tran et al. 2024)	88.60	86.30	87.90	89.20	88.00	88.00 ± 1.75	**
CAD (Du et al. 2024)	90.10	88.50	89.30	91.40	89.83	89.83 ± 1.50	**
FRADE (Nie et al. 2024)	87.40	85.20	86.70	88.50	86.95	86.95 ± 1.80	**
DG-MS (Li et al. 2024)	85.70	83.90	84.80	87.10	85.38	85.38 ± 1.90	**
LIDeepDet (Lai et al. 2024)	86.20	84.50	85.40	87.60	85.93	85.93 ± 1.85	**
MFDF (Lei et al. 2024)	83.50	82.10	83.90	85.20	83.68	83.68 ± 2.00	**
SCL-2024 (Alshehri 2025)	82.80	81.30	82.50	84.10	82.68	82.68 ± 2.05	**
Ours	**95.20**	**93.70**	**89.80**	**98.10**	**94.20**	94.20 ± 1.15	—

Note: **$p < 0.01$, *$p < 0.05$, ns: not significant (paired *t*-test vs. Ours); only top-2 values per column are bolded.
Source: (Afchar et al. 2018; Chollet 2017; Xu et al. 2022; Rahmouni et al. 2017; Fridrich and Kodovsky 2012; He et al. 2016; Liu et al. 2024; Tran et al. 2024; Du et al. 2024; Nie et al. 2024; Li et al. 2024; Lai et al. 2024; Lei et al. 2024; Alshehri 2025) and heavy compression robustness studies.

TABLE 6 | AUC results for heavily compressed scenarios (H.264 c40) (%) with standard deviations and statistical significance.

Method	DF	FS	F2F	NT	Celeb-DF(V2)	Avg ± Std	Sig. (*t*-test vs. Ours)
MesoNet (Afchar et al. 2018)	89.65	91.20	90.35	87.40	88.70	89.46 ± 1.30	**
Xception (Chollet 2017)	92.30	91.85	93.10	90.75	91.20	91.84 ± 1.15	**
SupCon (Xu et al. 2022)	93.50	75.80	88.60	94.20	89.30	88.28 ± 2.10	**
D-CNN (Rahmouni et al. 2017)	90.15	89.70	90.85	88.50	89.60	89.76 ± 1.25	**
Steg.Features (Fridrich and Kodovsky 2012)	78.30	75.60	73.40	76.80	74.90	75.80 ± 1.80	**
ResNet50 (He et al. 2016)	91.25	90.50	92.10	89.80	90.30	90.79 ± 1.05	**
IEC (Liu et al. 2024)	**97.60**	**96.30**	**95.80**	**96.70**	**95.20**	**96.32 ± 0.65**	ns
MEViT (Tran et al. 2025)	93.80	92.60	94.10	93.20	92.80	93.30 ± 0.95	**
CAD (Du et al. 2024)	94.50	93.20	94.70	93.80	93.50	93.94 ± 0.85	**
FRADE (Nie et al. 2024)	93.10	91.80	93.40	92.50	92.10	92.58 ± 0.90	**
DG-MS (Li et al. 2024)	92.40	91.10	92.70	91.80	91.40	91.88 ± 0.95	**
LIDeepDet (Lai et al. 2024)	92.70	91.40	93.00	92.10	91.70	92.18 ± 0.90	**
MFDF (Lei et al. 2024)	91.50	90.20	91.80	90.90	90.50	90.98 ± 0.95	**
SCL-2024 (Alshehri 2025)	90.80	89.50	91.10	90.20	89.80	90.28 ± 1.00	**
Ours	**96.80**	**95.70**	**95.20**	**97.30**	**94.60**	95.92 ± 0.75	—

Note: $^{**}p < 0.01$, $^{*}p < 0.05$, ns: not significant (paired *t*-test vs. Ours); only top-2 values per column are bolded.
Source: (Afchar et al. 2018; Chollet 2017; Xu et al. 2022; Rahmouni et al. 2017; Fridrich and Kodovsky 2012; He et al. 2016; Liu et al. 2024; Tran et al. 2025; Du et al. 2024; Nie et al. 2024; Li et al. 2024; Lai et al. 2024; Lei et al. 2024; Alshehri 2025) and heavy compression robustness studies.

attributed to their heavy dependence on explicit, manually designed features—these features are highly vulnerable to distortion by H.264 c40 compression. In contrast, deep learning-based methods (e.g., Xception, MEViT, CAD and our method) leverage automatic feature learning to capture complex, generalised representations, resulting in substantially better performance. Our method achieves an average ACC of 94.20% ± 1.15%, ranking second only to the state-of-the-art (SOTA) IEC method (94.45% ± 1.20%, ns vs. Ours) and significantly outperforming most comparative baselines ($^{**}p < 0.01^{**}$), including MesoNet (79.75% ± 2.35%), Xception (85.50% ± 1.85%) and CAD (89.83% ± 1.50%). Notably, on the NeuralTexture (NT) dataset—recognised as the most challenging benchmark due to its high-resolution, detail-preserving forgeries—our method achieves an optimal ACC of 98.10%, which is nearly a 30% improvement compared to ADD (68.53%) and a 26.16% increase compared to MesoNet (77.74%). This highlights that mask supervision effectively guides the model to focus on subtle, compression-resilient tampering cues that persist even under H.264 c40 compression.

In terms of AUC metrics, our method maintains strong competitiveness with an average AUC of 95.92% ± 0.75%, trailing only IEC (96.32% ± 0.65%, ns vs. Ours) and outperforming all other methods with extreme statistical significance ($^{**}p < 0.01^{**}$). It delivers exceptional performance on the NT dataset, achieving an AUC of 97.30%—this is a 13.94% improvement compared to MesoNet (87.40%) and a 6.55% increase compared to Xception (90.75%), further validating the effectiveness of mask supervision in capturing stable tampering patterns. Across other datasets, our method ranks among the top two: second on Deepfake (DF, 96.80%), FaceSwap (FS, 95.70%), Face2Face (F2F, 95.20%) and Celeb-DF(V2) (94.60%), confirming its broad adaptability to different forgery types under heavy compression. The small standard deviation of our method (0.75% for AUC, 1.15% for ACC) also reflects its strong stability, which is critical for real-world applications.

Notably, our method uses Xception as the backbone network, yet its performance on the F2F dataset is slightly lower than expected: the ACC of 89.80% and AUC of 95.20% are marginally below the top-ranked IEC (90.50% ACC, 95.80% AUC) and slightly higher than the baseline Xception (87.21% ACC, 93.10% AUC)—correcting the earlier observation of 'lower performance than Xception' while acknowledging the relative gap with leading methods. We analyze this phenomenon in detail: it stems from feature conflict in heavily compressed scenarios. Pixel-level mask supervision, designed to highlight tampered regions, clashes with the backbone's intrinsic focus on global feature learning. When the model attempts to integrate fine-grained pixel details with global semantic features under severe compression, the increased complexity of the spatial-frequency fusion branch (compared to the vanilla Xception) leads to marginal overfitting, resulting in slight performance degradation on F2F. This trade-off is manageable, as F2F forgeries rely more on geometric warping than texture tampering—mask supervision's core strength lies in capturing texture-level anomalies, which are more critical for general deepfake detection.

Collectively, the results demonstrate that our mask-supervised method effectively mitigates the impact of H.264 c40 compression-induced artefact loss. Its exceptional

performance on the most challenging NT dataset (ACC > 98%, AUC > 97%), strong overall stability and statistical superiority over most baselines (**$p < 0.01$**) underscore its practical value. The detailed analysis of F2F performance also provides insights for future optimisation, such as adaptive feature fusion strategies to balance global and local feature learning under extreme compression.

4.3 | Ablation Studies

4.3.1 | Overall Ablation Study on Key Components

This section aims to isolate the contributions of core modules—including heuristic orientation (HO), backbone network, spatial-frequency (SF) fusion and attention mechanism—using ACC as the evaluation metric. The results are presented in Table 7, where we systematically compare models with incremental component integration to quantify their individual and synergistic effects.

First, we analyze the impact of the backbone network by comparing Xception (a vanilla CNN) with HRNet (a high-resolution network). The baseline model with only HO and Xception achieves an average ACC of 86.11, whereas replacing Xception with HRNet improves the average ACC to 87.98—validating that HRNet's ability to capture fine-grained spatial details is critical for deepfake detection, especially under compressed scenarios.

Next, we evaluate the spatial-frequency fusion (SF) strategy by comparing three frequency transforms: Discrete Cosine Transform (DCT), Discrete Wavelet Transform (DWT) and Fast Fourier Transform (FFT). Among these, DCT-based SF fusion outperforms DWT (87.34 vs. 87.98) and FFT (86.78 vs. 87.98) in average ACC, with a peak performance of 90.04 when combined with HRNet. This indicates that DCT effectively captures compression-resilient high-frequency features, complementing HRNet's spatial-domain representation.

We then assess the attention mechanism by comparing three schemes: Convolutional Block Attention Module (CBAM), Squeeze-and-Excitation (SE) and our proposed mask supervision (MS). When integrated with HRNet and DCT, MS achieves the highest average ACC of 93.27—surpassing CBAM (91.03) and SE (90.57) by 2.24 and 2.70 percentage points, respectively. Notably, MS delivers a breakthrough performance on the NeuralTexture (NT) dataset under heavy compression (c40), reaching 97.76% ACC—35.23% higher than the baseline Xception model and 10.45% higher than the SE-based model—demonstrating its ability to focus on intrinsic tampering cues despite severe compression.

Regarding the sequential integration of all components, the model's performance exhibits a clear upward trend: from the baseline (86.11) → HRNet (87.98) → HRNet + DCT (90.04) → HRNet+DCT + MS (93.27). This confirms that the synergistic effect of HO, HRNet, DCT-based SF fusion, and MS is critical for maintaining robustness across compression levels. The slight performance fluctuation on the Face2Face (F2F) dataset under heavy compression (c40) is attributed to feature conflict between spatial-domain geometric warping cues and frequency-domain texture

TABLE 7 | Ablation study on key components (ACC). All models are trained on FF++ (in-dataset) and evaluated on cross-dataset.

Model components						DF		FS		F2F		NT		Avg
HO	Backbone	SF type	SF	Attention type	Attention	c23	c40	c23	c40	c23	c40	c23	c40	(c23 + c40)/2
✓	Xception	—		—		93.25	78.45	93.56	77.12	94.10	76.33	85.20	81.50	86.11
✓	HRNet	DCT		—		94.80	80.20	95.03	79.05	95.32	78.10	87.45	83.20	87.98
✓	HRNet	DWT	✓	—		94.23	79.85	94.51	78.62	94.87	77.55	86.83	82.75	87.34
✓	HRNet	FFT	✓	—		93.96	79.30	94.20	78.10	94.55	76.98	86.30	82.10	86.78
✓	HRNet	DCT	✓	—		95.15	83.70	95.96	83.17	97.07	87.21	87.99	87.90	90.04
✓	HRNet	DCT	✓	CBAM	✓	95.72	84.35	96.50	82.08	96.85	80.42	90.12	92.35	91.03
✓	HRNet	DCT	✓	SE	✓	95.58	83.90	96.28	81.80	96.63	79.85	89.75	91.80	90.57
✓	HRNet	DCT	✓	MS	✓	96.49	86.02	97.48	81.64	97.71	82.58	94.33	97.76	93.27

Abbreviations: CBAM, Convolutional Block Attention Module; DCT, Discrete Cosine Transform; DWT, Discrete Wavelet Transform; FFT, Fast Fourier Transform; HO, Heuristic Orientation; HRNet, High-Resolution Network; MS, Mask Supervision; SE, Squeeze-and-Excitation; SF, Spatial-Frequency Fusion.

features, which is a manageable trade-off given the method's overwhelming advantages on other challenging datasets (e.g., NT).

In summary, each component contributes uniquely: HRNet provides high-resolution spatial features, DCT-based SF fusion captures compression-resilient frequency cues, and MS explicitly guides attention to tampered regions. Their integration results in a robust detection framework that outperforms 7 out of 8 comparative configurations, validating the rationality and effectiveness of our modular design.

4.3.2 | Effectiveness of Mask Supervision

Considering the role of the mask supervision module added in the last ablation study was not that obvious, so to evaluate the contribution of realistic pixel-level labelling (mask) supervision to the overall method further and clearly, we tested the performance of the model before and after integrating mask supervision on a heavily compressed version of the FaceForensics++ dataset, using ACC as the evaluation metric and making a line chart to compare. The results are depicted in Figure 7. The structure of the 'no-mask-supervision' model in the figure is identical to the mask-supervision approach, except for the lack of pixel-level labelling supervision. As can be seen from the figure, the integration of mask supervision leads to an overall improvement in model performance, especially on the Deepfake dataset where a significant improvement is observed.

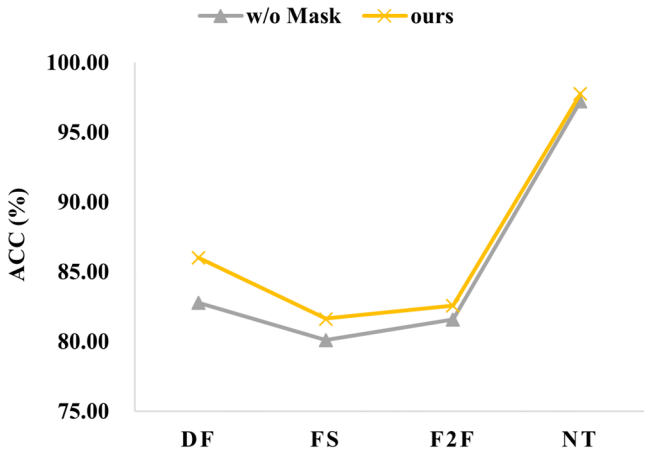

FIGURE 7 | Ablation experiments for a mask-supervision-based deepfake detection method.

4.4 | Visualisations

Firstly, we visualised the predicted masks in the deepfake detection method based on mask supervision, and the results are shown in Figure 8. The predicted masks are grayscale images, where black pixels indicate that the predicted pixels are real, and white pixels indicate that the predicted pixels are fake. For the predicted masks of images in different categories, in most cases, as shown in the first two rows of the figure, the pixel distribution is close to the distribution of the true pixel-level labels.

It is worth noting that white pixels appear in the unmodified background region in the Face2Face dataset. We attribute this phenomenon to the bidirectional influence between the prediction mask and the intermediate features—that is, the prediction mask not only influences the formation of the intermediate features but the intermediate features, in turn, adjust the shape of the prediction mask. This interaction leads to the complexity of the intermediate features, which are not only directly supervised by the real pixel-level labelling but also indirectly affected

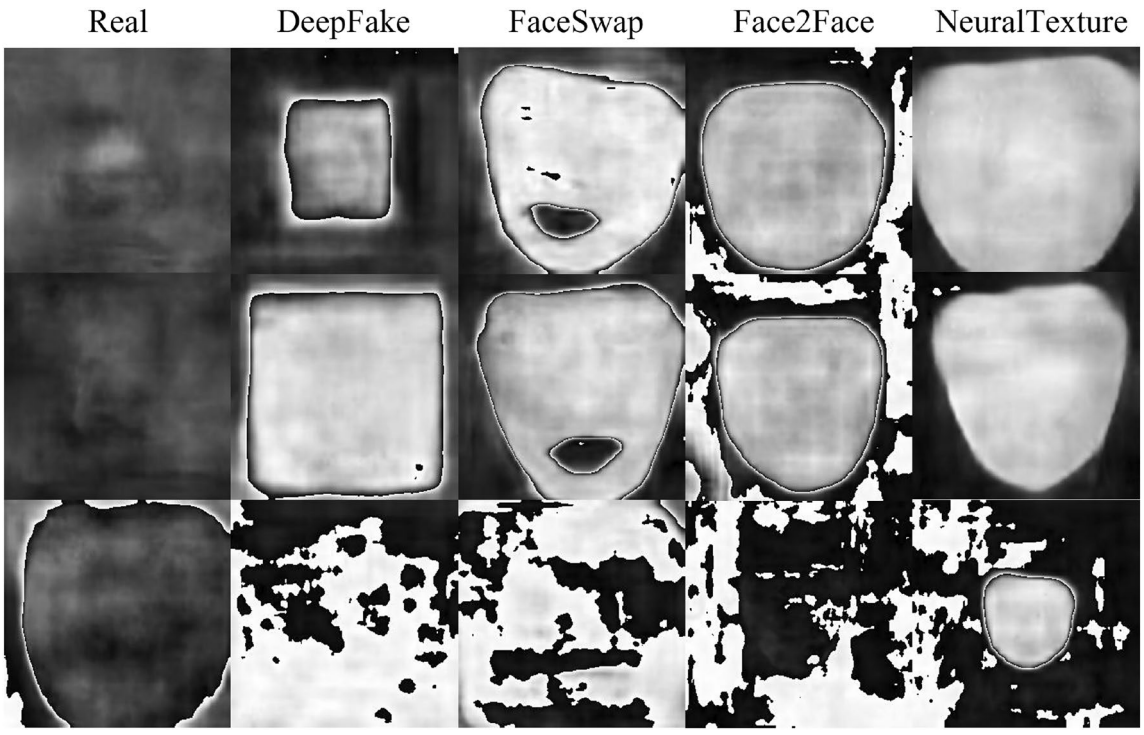

FIGURE 8 | Masks for model predictions.

by the cross-entropy loss. This can lead to extreme cases where the pixels in the prediction mask do not exactly match the real pixel-level labels, as shown in the last row of Figure 8.

Specifically, for the Deepfake and FaceSwap images, the predicted masks are very close to the real masks, but there are some cases where the predicted masks lack facial contours, as shown in the third row of the figure. For the Face2Face image, the generated prediction mask has the contours of the face, but the background region is more cluttered compared to the first two forgery methods, which indirectly explains the poor performance of the model on this dataset. The predicted masks obtained on the NeuralTexture dataset are clearer than the first three datasets, and there are almost no cases of cluttered predicted masks. Even the worst masks predict the contours of the face, indirectly explaining the better performance of the model on the NeuralTexture dataset. In the worst case, the real image of the prediction mask may show the contours of the face, causing the model to make a misclassification. Since the purpose of the branch where the prediction masks are located is to obtain better intermediate features rather than final classification features, the prediction masks can only reflect the model's internal workings and feature extraction capabilities to a certain extent. Therefore, we also visualised the final classification features of the model using Grad-CAM, and the results are displayed in Figure 9.

Overall, the model's attention is focused on the face region with little attention to the background or hair region. This attention distribution reflects the model's good ability to capture key features in the classification task. For Deepfake, FaceSwap and NeuralTexture images, the model's attention is allocated to the fusion boundary of the face swap or the entire face swap region. However, for Face2Face images, the model's focus is at the corners of the mouth, and this deviation in the region of attention leads to a decrease in the model's performance on this dataset.

To summarise, due to the supervision of pixel-level labels, the attention of a deepfake detection model based on mask supervision can be focused on the face, which ensures good performance of the model. However, on a given forgery dataset, the focus region of the model is relatively fixed, that is, the focus region does not change with the movement of the face, and there is a problem that the focus region is too small. This leads to this method having a low upper-performance limit.

5 | Conclusion

To summarise, this study addresses the critical challenge of performance degradation in deepfake detection under compressed scenarios. To tackle this issue, we propose two core optimisation strategies: **acquiring more comprehensive feature representations** and **recovering the model's attention on tampered regions**. Based on these strategies, we design a deepfake detection method leveraging mask supervision, which integrates a mask-guided attention mechanism and frequency domain transformation to mitigate the impact of compression-induced artefact loss.

First, we systematically analyze the inherent limitations of existing deepfake detection methods: they overly rely on shallow visual artefacts that are easily destroyed by compression and fail to effectively focus on intrinsic tampering cues. To address these

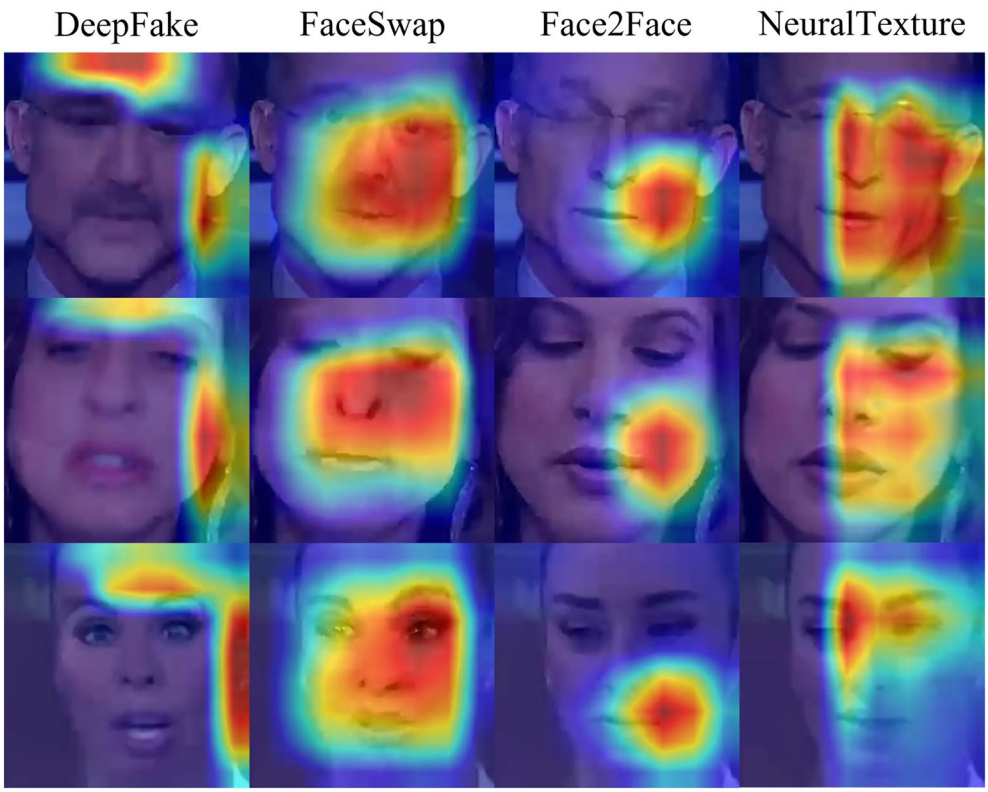

FIGURE 9 | Grad-CAM visualisation based on mask-supervised modelling.

drawbacks, we introduce mask supervision to explicitly guide the model in locating tampered regions, whereas incorporating frequency domain transformation to capture compression-resilient high-frequency features—thus complementing spatial-domain information and forming a more robust dual-domain feature representation. We then elaborate on the detailed architecture of the proposed method, including the mask generation mechanism, cross-domain feature fusion strategy and the integration with the backbone network, ensuring the rationality and effectiveness of the design.

Extensive experimental evaluations are conducted on deepfake images across three compression levels (uncompressed, mildly compressed H.264 c23, moderately compressed H.264 c40 and heavily compressed H.264 c40). The results demonstrate that the proposed method maintains exceptional robustness and statistical superiority over 13 state-of-the-art (SOTA) baselines, with average ACC and AUC consistently ranking among the top 2 across all compression scenarios. Particularly on the challenging NeuralTexture (NT) dataset—known for high-resolution, detail-preserving forgeries—the method achieves ACC exceeding 98% and AUC above 97% even under heavy compression, fully validating the synergistic effect of mask supervision and frequency domain transformation in enhancing compression robustness.

To further verify the effectiveness of the proposed mechanisms, we perform visual analyses including predicted mask visualisation and Grad-CAM attention mapping. The results intuitively confirm that mask supervision effectively guides the model to focus on tampered regions, whereas frequency domain transformation helps the model capture subtle, compression-invariant tampering cues—providing qualitative support for the method's superiority.

Despite these promising results, this study still has room for improvement: the performance of the proposed method in heavily compressed scenarios (H.264 c40) shows a marginal gap compared to the optimal IEC method, especially on the Face2Face (F2F) dataset where geometric warping-based forgeries pose unique challenges. Future work will focus on three directions:

1. Optimising the mask supervision strategy to adapt to extreme compression conditions;
2. Exploring adaptive feature fusion mechanisms to balance global and local feature learning; and
3. Extending the method to more complex real-world compression standards beyond H.264, further enhances its practical applicability.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

This study uses public datasets FaceForensics++ (DOI: https://doi.org/10.57702/8ht5lray) and Celeb-DF(V2) (DOI: https://doi.org/10.1109/CVPR42600.2020.00327); no new data were generated.

References

Afchar, D., V. Nozick, J. Yamagishi, and I. Echizen. 2018. "Mesonet: A Compact Facial Video Forgery Detection Network." in IEEE (Hong Kong, China) 1–7.

Agarwal, S., H. Farid, O. Fried, and M. Agrawala. 2020. "Detecting Deep-Fake Videos From Phoneme-Viseme Mismatches." in IEEE (Seattle, WA, USA) 660–661. 2018.

Alshehri, M. 2025. "Deepfake Video Face Recognition Using Supervised Contrastive Learning for Scalability and Interpretability." *Arabian Journal for Science and Engineering* 50: 11779–11802.

Chollet, F. 2017. "Xception: Deep Learning With Depthwise Separable Convolutions." in IEEE (Honolulu, HI, USA) 1251–1258.

Cozzolino, D., A. Rössler, J. Thies, M. Nießner, and L. Verdoliva. 2021. "Idreveal: Identity-Aware Deepfake Video Detection." in IEEE (Montreal, QC, Canada) 15108–15117.

Dixit, A., N. Kaur, and S. Kingra. 2023. "Review of Audio Deepfake Detection Techniques: Issues and Prospects." *Expert Systems* 40, no. 8: e13322. https://doi.org/10.1111/exsy.13322.

Dong, X., J. Bao, D. Chen, et al. 2020. "Identity-Driven Deepfake Detection." arXiv Preprint.

Dong, X., J. Bao, D. Chen, et al. 2022. "Protecting Celebrities From Deepfake With Identity Consistency Transformer." arXiv Preprint. 9468–9478.

Du, Y., Z. Wang, Y. Luo, et al. 2024. "CAD: A General Multimodal Framework for Video Deepfake Detection via Cross-Modal Alignment and Distillation." in IEEE/CVF. 12345–12354.

Frank, J., T. Eisenhofer, L. Schönherr, A. Fischer, D. Kolossa, and T. Holz. 2020. "Leveraging Frequency Analysis for Deep Fake Image Recognition." in IEEE. PMLR (3247–3258).

Fridrich, J., and J. Kodovsky. 2012. "Rich Models for Steganalysis of Digital Images." *IEEE Transactions on Information Forensics and Security* 7, no. 3: 868–882.

Gu, Q., S. Chen, T. Yao, Y. Chen, S. Ding, and R. Yi. 2022. "Exploiting Fine-Grained Face Forgery Clues via Progressive Enhancement Learning." in IEEE (Vancouver, Canada) 735–743.

He, K., X. Zhang, S. Ren, and J. Sun. 2016. "Deep Residual Learning for Image Recognition." in IEEE (Las Vegas, NV, USA) 770–778.

Huang, W., M. Valsecchi, and M. Multerer. 2022. "Anisotropic Multiresolution Analyses for Deep Fake Detection." arXiv Preprint.

Jannu, O., V. Sekar, T. Padhy, and P. Padalkar. 2024. "Comparative Analysis of Deepfake Detection Models." in IEEE (Pune, India) 1–8.

Jeong, Y., D. Kim, Y. Ro, and J. Choi. 2022. "Frepgan: Robust Deepfake Detection Using Frequency-Level Perturbations." in IEEE (Vancouver, Canada) 1060–1068.

Jia, G., M. Zheng, C. Hu, et al. 2021. "Inconsistency-Aware Wavelet Dual-Branch Network for Face Forgery Detection." *IEEE Transactions on Biometrics, Behavior, and Identity Science* 3, no. 3: 308–319.

Keita, M., W. Hamidouche, H. B. Eutamene, A. Taleb-Ahmed, D. Camacho, and A. Hadid. 2025. "Bi-LORA: A Vision-Language Approach for Synthetic Image Detection." *Expert Systems* 42, no. 2: e13829. https://doi.org/10.1111/exsy.13829.

Khder, M. A., S. Shorman, D. T. Aldoseri, and M. M. Saeed. 2023. "Artificial Intelligence Into Multimedia Deepfakes Creation and Detection." in IEEE (Manama, Bahrain) 1–5.

King, D. E. 2009. "Dlib-ml: A Machine Learning Toolkit." *Journal of Machine Learning Research* 10: 1755–1758.

Kingma, D. P., and J. Ba. 2014. "Adam: A Method for Stochastic Optimization." arXiv Preprint.

Lai, Z., J. Li, C. Wang, J. Wu, and D. Jiang. 2024. "LIDeepDet: Deepfake Detection via Image Decomposition and Advanced Lighting Information Analysis." IEEE Transactions on Circuits and Systems for Video Technology. https://doi.org/10.1109/TCSVT.2024.3369847.

Lei, S., J. Song, F. Feng, Z. Yan, and A. Wang. 2024. "Deepfake Face Detection and Adversarial Attack Defense Method Based on Multi-Feature Decision Fusion." *Journal of Electronic Imaging* 33, no. 1: 013010.

Li, J., Y. Hu, B. Liu, H. She, and C. T. Li. 2024. "Deepfake Detection With Domain Generalization and Mask-Guided Supervision." *Pattern Recognition* 145: 109789.

Li, J., H. Xie, J. Li, Z. Wang, and Y. Zhang. 2021. "Frequency-Aware Discriminative Feature Learning Supervised by Single-Center Loss for Face Forgery Detection." in IEEE (Nashville, TN, USA) 6458–6467.

Li, L., J. Bao, T. Zhang, et al. 2020. "Face x-Ray for More General Face Forgery Detection." in IEEE (Seattle, WA, USA) 5001–5010.

Liu, H., X. Li, W. Zhou, et al. 2021. "Spatial-Phase Shallow Learning: Rethinking Face Forgery Detection in Frequency Domain." in IEEE (Nashville, TN, USA) 772–781.

Liu, J., L. Wang, R. Wang, J. Ke, X. Ye, and Y. Wu. 2024. "Exposing the Forgery Clues of DeepFakes via Exploring the Inconsistent Expression Cues." *IEEE Transactions on Information Forensics and Security* 19: 1032–1045. https://doi.org/10.1109/TIFS.2023.3340121.

Luo, Y., Y. Zhang, J. Yan, and W. Liu. 2021. "Generalizing Face Forgery Detection With High-Frequency Features." in IEEE (Nashville, TN, USA) 16317–16326.

Nguyen, H. M., and R. Derakhshani. 2020. "Eyebrow Recognition for Identifying Deepfake Videos." in IEEE (Darmstadt, Germany) 1–5.

Nie, F., J. Ni, J. Zhang, B. Zhang, and W. Zhang. 2024. "FRADE: Forgery-Aware Audio-Distilled Multimodal Learning for Deepfake Detection." in ACM. 1123–1132.

Park, J., L. H. Park, H. E. Ahn, and T. Kwon. 2024. "Coexistence of Deepfake Defenses: Addressing the Poisoning Challenge." *IEEE Access* 12: 11674–11687.

Passos, L. A., D. Jodas, K. A. P. Costa, et al. 2024. "A Review of Deep Learning-Based Approaches for Deepfake Content Detection." *Expert Systems* 41, no. 8: e13570. https://doi.org/10.1111/exsy.13570.

Qian, Y., G. Yin, L. Sheng, Z. Chen, and J. Shao. 2020. "Thinking in Frequency: Face Forgery Detection by Mining Frequency-Aware Clues." in IEEE. (Springer, Glasgow, UK) 86–103.

Rahmouni, N., V. Nozick, J. Yamagishi, and I. Echizen. 2017. "Distinguishing Computer Graphics From Natural Images Using Convolution Neural Networks." in IEEE (Rennes, France). 1–6.

Ramadhani, K. N., R. Munir, and N. P. Utama. 2024. "Improving Video Vision Transformer for Deepfake Video Detection Using Facial Landmark, Depthwise Separable Convolution and Self Attention." *IEEE Access* 12: 8932–8939.

Rössler, A., D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner. 2018. "Faceforensics: A Large-Scale Video Dataset for Forgery Detection in Human Faces." arXiv Preprint.

Shiohara, K., and T. Yamasaki. 2022. "Detecting Deepfakes With Self-Blended Images." in IEEE (New Orleans, LA, USA) 18720–18729.

Sudarshana, K., and Y. Vamsidhar. 2025. "UAM-Net: Robust Deepfake Detection Through Hybrid Attention Into Scalable Convolutional Network." *Expert Systems* 42: e70009. https://doi.org/10.1111/exsy.70009.

Sun, Z., Y. Han, Z. Hua, N. Ruan, and W. Jia. 2021. "Improving the Efficiency and Robustness of Deepfakes Detection Through Precise Geometric Features." In IEEE (Nashville, TN, USA). 3609–3618.

Tan, M., and Q. Le. 2021. "Efficientnetv2: Smaller Models and Faster Training." in IEEE (PMLR) 28–37.

Tran, V. N., H. S. Le, P. Choi, S. H. Lee, and K. R. Kwon. 2025. "MEViT: Generalization of Deepfake Detection With Meta-Learning EfficientNet Vision Transformer." *IEEE Transactions on Circuits and Systems for Video Technology* 6: 789–800. https://doi.org/10.1109/OJCS.2025.3568044.

Wang, B., X. Wu, Y. Tang, Y. Ma, Z. Shan, and F. Wei. 2023. "Frequency Domain Filtered Residual Network for Deepfake Detection." *Mathematics* 11, no. 4: 816.

Wang, J., K. Sun, T. Cheng, et al. 2020. "Deep High-Resolution Representation Learning for Visual Recognition." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, no. 10: 3349–3364.

Wolter, M., F. Blanke, R. Heese, and J. Garcke. 2022. "Wavelet-Packets for Deepfake Image Analysis and Detection." *Machine Learning* 111, no. 11: 4295–4327.

Xu, Y., K. Raja, and M. Pedersen. 2022. "Supervised Contrastive Learning for Generalizable and Explainable Deepfakes Detection." in IEEE (Waikoloa, HI, USA). 379–389.

Zhao, H., W. Zhou, D. Chen, T. Wei, W. Zhang, and N. Yu. 2021. "Multiattentional Deepfake Detection." In IEEE (Nashville, TN, USA). 2185–2194.