Waseem *et al.*
EURASIP Journal on Image and Video Processing (2023) 2023:14
https://doi.org/10.1186/s13640-023-00614-z

EURASIP Journal on Image and Video Processing

RESEARCH Open Access

Multi-attention-based approach for deepfake face and expression swap detection and localization

Saima Waseem[1*], Syed Abdul Rahman Syed Abu-Bakar[1], Zaid Omar[1], Bilal Ashfaq Ahmed[2], Saba Baloch[1] and Adel Hafeezallah[3]

*Correspondence: syed@fke.utm.my

[1] Faculty of Electrical Engineering, Universiti Teknologi Malaysia, Johor, Malaysia
[2] Faculty of Computing, Universiti Teknologi Malaysia, Johor, Malaysia
[3] Department of Electrical Engineering, Taibah University, Madinah, Saudi Arabia

Abstract

Advancements in facial manipulation technology have resulted in highly realistic and indistinguishable face and expression swap videos. However, this has also raised concerns regarding the security risks associated with deepfakes. In the field of multimedia forensics, the detection and precise localization of image forgery has become essential tasks. Current deepfake detectors perform well with high-quality faces within specific datasets, but often struggle to maintain their performance when evaluated across different datasets. To this end, we propose an attention-based multi-task approach to improve feature maps for classification and localization tasks. The encoder and the attention-based decoder of our network generate localized maps that highlight regions with information about the type of manipulation. These localized features are shared with the classification network, improving its performance. Instead of using encoded spatial features, attention-based localized features from the decoder's first layer are combined with frequency domain features to create a discriminative representation for deepfake detection. Through extensive experiments on face and expression swap datasets, we demonstrate that our method achieves competitive performance in comparison to state-of-the-art deepfake detection approaches in both in-dataset and cross-dataset scenarios. Code is available at https://github.com/saimawaseem/Multi-Attention-Based-Approach-for-Deepfake-Face-and-Expression-Swap-Detection-and-Localization.

Keywords: Deepfake, Face-forgery, Face-swap, Re-enactment, Forensics

1 Introduction

Deepfake techniques have recently achieved significant success due to advances in generative models [1–5]. These techniques empower individuals with the ability to manipulate facial features within an image, resulting in the creation of forged faces. The current approaches have the capability to generate high-quality fake content that appears indistinguishable from real media to the human eye. Numerous instances of deepfake have been exploited, particularly in politics and pornography [6, 7]. This misinformation has caused people to worry about fraud and credibility issues in society. Face (identity) and

expression swap are two well-known forms of deepfake face manipulation. Expression-swap or re-enactment techniques enable the transfer of expressions from one person to another while keeping the original subject's identity unchanged. In contrast, identity or face swap involves replacing the face of one person with the face of another individual [8]. A well-designed facial expression can effectively convince others to agree with someone's perspective without any verbal communication, and with a deepfake face swap, it becomes possible to portray an individual's physical presence in a particular location where they were not actually present. To effectively combat these deepfakes, the development of robust and reliable face forgery forensics is important to ensure the integrity and ethical standards of multimedia content.

Existing deepfake detection techniques typically frame deepfake as a binary classification task. These approaches heavily rely on deep neural networks (DNN) [9–17]. Nevertheless, some researchers have explored alternative techniques [18–21] utilizing hand-crafted features for deepfake detection. However, with the rapid development of deepfake synthesis techniques [4, 22, 23], the performances of hand-crafted approaches are not satisfactory [8]. A common approach among DNN methods involves extracting video frames and using a convolution neural network (CNN) with a fully connected layer for classification. However, these methods overlook correlations between distant positions by focusing on information within each receptive field. As a result, they rely on superficial correlations to differentiate between real and manipulated images. Due to the independent and evenly distributed training-test split, these simplified patterns have a random probability of being effective on unseen test sets, making them susceptible to overfitting. Consequently, their effectiveness is limited to the manipulation methods they were explicitly trained on, and these approaches exhibit significant performance decline when detecting unseen face manipulations. To address this limitation, recent deepfake detection algorithms have incorporated the concept of the attention mechanism into CNNs [24] to enhance both within-dataset and cross-dataset performance by expanding the areas of local image features. Different manipulation methods, such as face swap and expression swap, have unique characteristics and patterns of manipulation, as shown in Fig. 1. These variations in forgery patterns pose a challenge in maintaining similarity among each manipulation method, which may result in overfitting and a decrease in overall performance [25].

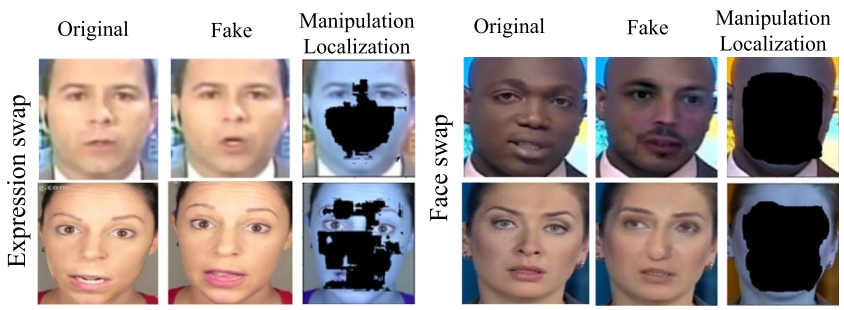

Fig. 1 Illustration of deepfake face swap and expression swap manipulations along with the corresponding localization maps generated by the proposed approach. The localization maps emphasize the specific regions on the face that have undergone manipulation

Recent deepfake generation techniques, like GANs, often employ encoder–decoder architectures in their generators. The decoder incorporates an upsampling design to enlarge the feature maps generated by the encoder, resulting in a colorful image. However, this upsampling process hinders GAN models from accurately reproducing the spectral distributions of real training data [13, 26]. Consequently, fake images exhibit distinct artifacts in their frequency spectrum, which can be exploited to differentiate them from real images [13]. These frequency-related artifacts are commonly observed in various deepfake manipulations, especially in scenarios that involve compression where spatial information is significantly degraded [27].

We hypothesize that by appropriately assessing an image's spatial and spectral information, the network can effectively focus on critical regions for decision-making. Here, we propose an attention-based multi-task learning technique that effectively integrates spatial and spectral information to classify the facial images as real or fake, while simultaneously localizing modified regions within the face, specifically in deepfake facial manipulation subcategories, i.e., face swap and expression swap (face re-enactment), depicted in Fig. 2. Accurate localization of manipulated regions is vital in multimedia forensics for a comprehensive understanding of deepfake forgeries, as high-resolution localization maps provide valuable insights into the specific type of manipulation employed. To address this, we introduce a simple attention-based learning technique to localize potential areas of manipulation. Explicitly localizing these manipulated regions through an attentional mechanism provides two benefits: it suppresses irrelevant information, directing the network's attention to manipulated areas, thereby avoiding disruptions and improving the network's understanding of modified regions.

Our experimental results demonstrate that the proposed attention-based manipulation localization and detection technique significantly improves performance in within-dataset and cross-dataset evaluations. Experimental results on popular deepfake datasets, such as FaceForensics++ [28], CelebDF [8], and DFDC-P [29] demonstrate the competitive performance of our approach compared to state-of-the-art methods. Our contributions can be summarized as follows:

- We present the image features learning scheme at local and global levels using a dual attention mechanism (spatial and channel) by jointly integrating convolutional encoder and decoder features to localize pixel-level image forgeries.

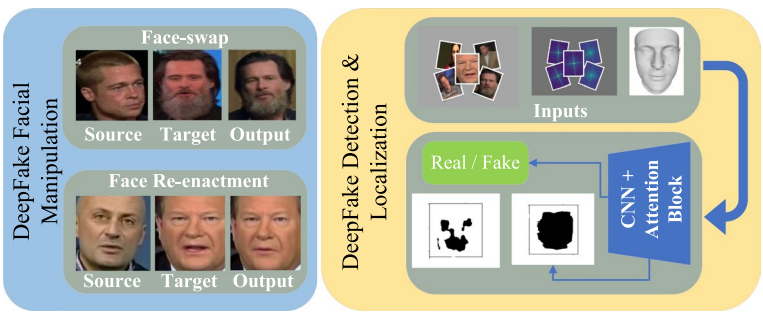

Fig. 2 Overview of deepfake detection problem

- Our proposed model demonstrates robustness for both cross-dataset and within-dataset evaluations by effectively combining frequency and localized spatial features.

2 Related work

This section provides a concise overview of prior research relevant to detecting and localizing deepfakes.

2.1 Manipulated detection

One common deepfake detection approach is to treat a video as a sequence of still images and perform operations on them. Various techniques have been explored, such as capturing unique low-level camera features to detect fake faces [30], estimating inconsistencies in head pose [19], and utilizing flaws in eye-blinking patterns and other facial features for deepfake classification [18, 31]. However, these methods are not effective in detecting advanced deepfake manipulation techniques. Several deep neural network-based solutions have been developed to differentiate between real and fake faces. These include MesoNet [12], Capsule Network [10], XceptionNet [32], EfficientNet [33], F^3 Net [16] and GocNet [17]. Various features, such as spatial, steganographic, and temporal features [14, 15, 34, 35], as well as frequency dependent cues [36], multi-scale Laplacian of Gaussian (LoG) operator [11], and motion features with a fine-grained weighting of inter-class distances [37] have been investigated for deepfake detection. Despite these efforts, challenges persist in detecting realistic deepfakes. Sun et al. [38] introduced Dual Contrastive Learning (DCL) approach to analyze real and fake paired data for deepfake detection. Multi-attention Deepfake Detection (MaDD) [24] presented a framework that captures artifacts using multiple attention maps. However, it lacks strong supervision and struggles to identify minor forgery traces in quality-degraded videos. Wodajo et al. [39] combined vision transformers with CNNs (CViT) to capture local and global features from face images, but at the cost of increased computational complexity due to a high number of parameters. Hua et al. [40] proposed an interpretable model for fake face detection by establishing a patch-channel correspondence that provides evidence for fake face detection. However, this approach faces limitations in quantifying the degree of interpretability and optimizing the patch-channel correspondence because of strong channel correlation and computational complexity.

2.2 Forgery localization

In addition to classification, certain techniques are specifically designed to focus on localizing the manipulated regions. Nguyen et al. [41] utilized a multi-task learning strategy with a Y-shaped architecture to simultaneously locate modified video regions and detect manipulation. Li et al. [42] presented an X-ray approach for faces to detect boundaries around manipulated face regions. However, this method relies on external training data and has lower performance when image quality varies, such as compression or blurring, which can affect the detection of boundary traces. Liu et al. [43] introduced an automated machine-learning approach for deepfake detection and localization, reducing the need for manual network design. Dang et al. [44] proposed supervised and

weakly supervised strategies for estimating image-specific attention maps to localize manipulated regions in face images. However, this approach is sensitive to compression. Therefore, it is crucial to prioritize robust localization of the manipulated regions to address the impact of compression. Our goal is to achieve consistent forgery localization even when image quality is compromised at different levels. Using pixel-level localization, our approach aims to improve the generalization performance of deepfake detection. Unlike previous methods focusing on spatial features, our approach simultaneously learns spatial features and frequency-related patterns.

By effectively incorporating spatial and spectral information, we developed a multi-task learning approach to classify facial images as real or fake and to localize modified regions in the face. For manipulation localization, we introduce an attention-based encoder–decoder architecture that integrates semantic information extraction. Our approach uses an attention-based U-Net architecture with frequency features in the detection stream, resulting in improved classification performance. To the best of our knowledge, this is the first study to use U-Net with a spatial and channel-specific attention mechanism for detecting and localizing face manipulations.

3 Proposed method

In contrast to single-objective approaches, our method utilizes attention-based localization and classification networks to generate the probability of an input image being forged or real and simultaneously provide localized maps highlighting manipulated regions within each input video frame, as shown in Fig. 3. The proposed model operates

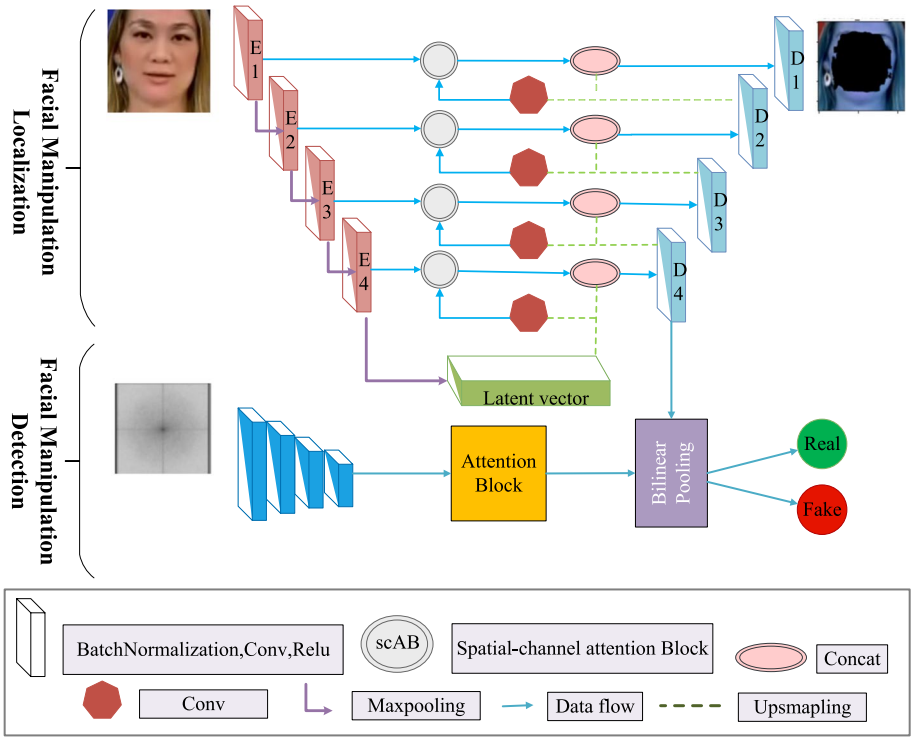

Fig. 3 Illustration of the pipeline used in the proposed method for detecting and localizing deepfake facial manipulation. Face image and its spectral (FFT) representation are used as input

on a tuple dataset denoted as $H = (A_i, B_i, D_i, y_i)_{i=1}^{N}$, where $A_i \in \mathbb{R}^{H\times W\times 3}$ represents a 2D image of a face, $B_i \in \mathbb{R}^{H\times W\times 1}$ corresponds to the reference input mask for each fake face type, which includes the face swap mask that covers the entire face area and the expression swap mask, representing the facial structure as shown in Fig. 6. D_i is the spectrum coefficient in the frequency domain, and $y_i \in (0, 1)$ serves as a label indicating whether the input A_i has been manipulated or not. The subscript i represents data points from the face and frequency spectrum dataset. Our main objective is to train a model to determine whether a test image has been manipulated and, if so, to what extent. Localizing the modifications in facial images requires focusing on the specific regions affected by each type of manipulation. Thus, for Facial Manipulation Localization (FML), we use a dataset $\{(A_i, B_i)\}_{i=1}^{N}$, while for Facial Manipulation Detection (FMD), we utilize a dataset of tuples $\{(D_i, y_i)\}_{i=1}^{N}$. To classify manipulated images by combining attention-based spatial and spectral information from two network streams, we replace Global Average Pooling with Bilinear Pooling (BP).

3.1 Facial manipulation localization (FML)

We propose Residual U-Net with spatial channel attention block (scAB) to focus on the face regions for deepfake localization during the learning process. In our approach, the encoder directly receives the input image, and during the decoding phase, both the encoder and the features from the preceding layer decoder undergo processing through the scAB to generate decoder features, as illustrated in Fig. 3. At each skip connection of Residual U-Net, we employ the scAB to dynamically learn the location and semantic information.

Given an image A_i passed through the network to produce encoder features $f_e \in \mathbb{R}^{C\times H\times W}$, and decoder features $f_d \in \mathbb{R}^{C\times H\times W}$, here, C, H, and W correspond to the feature map's channel count, height, and width, respectively. ScAB generates spatial attention map $\mathrm{SAM} \in \mathbb{R}^{1\times H\times W}$ and channel attention map $\mathrm{CAM} \in \mathbb{R}^{C\times 1\times 1}$ using the encoder f_e and decoder features f_d as depicted in Fig. 4. The spatial attention block (sAB) directs the model's attention toward relevant deep spatial structures. On the other hand, the channel attention block (cAB) acts as a bridge, closing the semantic gap between the encoder and decoder features by incorporating extra contextual information into the lower-level encoding features, thereby enhancing the overall understanding of the data.

The $\mathrm{SAM}(f_e, f_d)$ is computed by applying the Average Pooling and Max Pooling along the channel dimension of f_e and f_d. The resulting maps are summed-up and passed through a sigmoid function.

$$\mathrm{SAM}_e(f_e) = f_1^{\nu\times\nu}([(f_e)^s_{\mathrm{avg}}, (f_e)^s_{\mathrm{max}}]) \tag{1}$$

$$\mathrm{SAM}_d(f_d) = f_1^{\nu\times\nu}([(f_d)^s_{\mathrm{avg}}, (f_d)^s_{\mathrm{max}}]) \tag{2}$$

$$\mathrm{SAM}(f_e, f_d) = \sigma(\mathrm{SAM}_e(f_e), \mathrm{SAM}_d(f_d)) \tag{3}$$

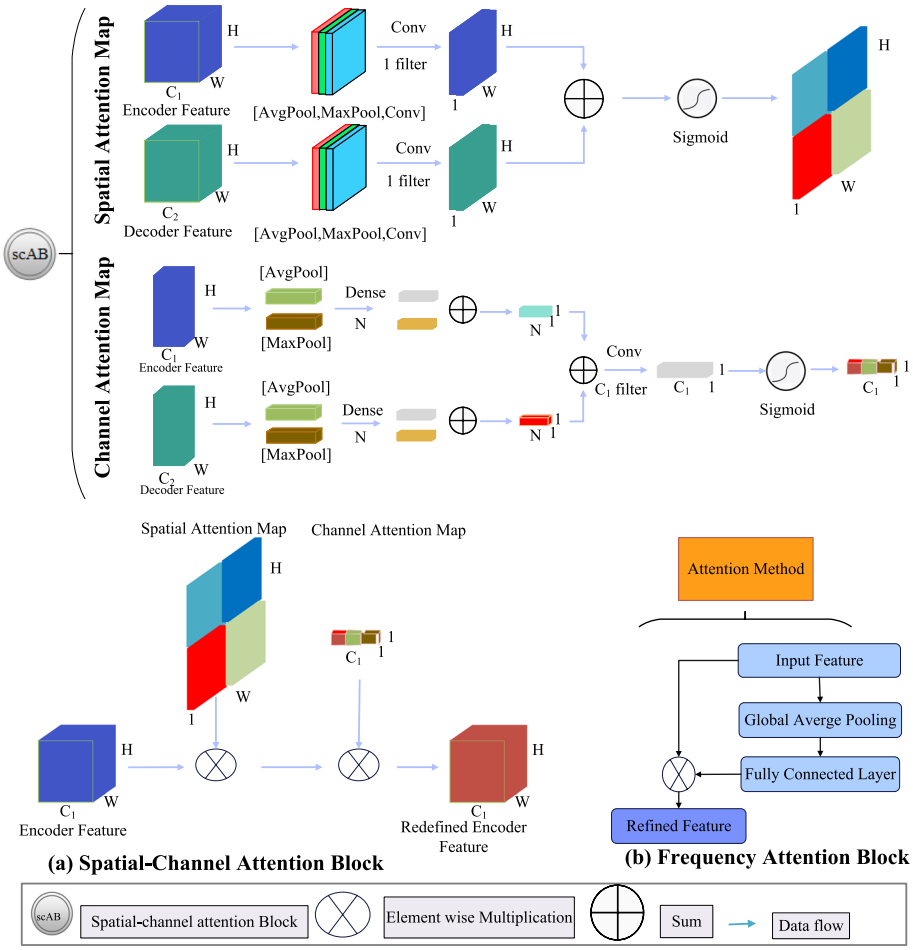

Fig. 4 Illustration of Spatial Channel Attention Block (scAB) and attention block for frequency spectrum features

The convolution operation is represented as $f_1^{\nu\times\nu}$ with a filter of 7×7 size, and the sigmoid function is denoted as σ. The low-level encoder features possess valuable spatial details but lack semantic information. Combining low-level encoders with high-level decoders without considering semantic differences can adversely affect the localization results. To improve fusion effectiveness, we integrate semantic concepts into low-level features using convolutional feature inter-dependencies. This is accomplished through the CAM technique, which facilitates meaningful feature discrimination [45].

To calculate CAM(f_e,f_d), we employ Average Pooling and Max Pooling techniques to reduce spatial information in both encoder and decoder features, as inspired by [46]. Subsequently, these compressed features are passed through N Dense layer with u units. It is essential to note that u varies for each Dense layer. The Dense layer operation is responsible for detecting channel dependencies and producing squeeze channel attention maps. The individual output attention maps $\text{CAM}_e(f_e)$ and $\text{CAM}_d(f_d)$ are then combined using element-wise summation. The resulting sum undergoes C_1 convolutions, followed by a sigmoid function, to obtain the final CAM representation. In summary, the computation of CAM(f_e,f_d) is as follows:

$$\mathrm{CAM}_e(f_e) = f_N^u((f_e)_{\mathrm{avg}}^c) + f_N^u((f_e)_{\mathrm{max}}^c), \tag{4}$$

$$\mathrm{CAM}_d(f_d) = f_N^u((f_d)_{\mathrm{avg}}^c) + f_N^u((f_d)_{\mathrm{max}}^c), \tag{5}$$

$$\mathrm{CAM}(f_e, f_d) = \sigma(F_{c_1}^{1\times 1}(\mathrm{CAM}_e(f_e) + \mathrm{CAM}_d(f_d))). \tag{6}$$

The input encoder layer features are enhanced by multiplying them with the scAB output, effectively incorporating the benefits of both SAM and CAM. This process is depicted in Fig. 4.

$$F_r = f_e \bigotimes \mathrm{SAM}(f_e, f_d) \bigotimes \mathrm{CAM}(f_e, f_d) \tag{7}$$

The element-wise multiplication operation $\otimes$ is applied to preserve the spatial and channel dimensions of the input feature map in both SAM and CAM. The refined features F_r, and the decoder features f_d are concatenated and passed to the convolutional layer to build the decoder features for the next layer.

3.2 Facial manipulation detection (FMD)

To improve image quality, common upsampling techniques are employed in autoencoders [47] or GANs [26]. These techniques increase the pixel dimensions vertically and horizontally by a factor of m, utilizing the low-resolution encoded image as input. By leveraging the property of Discrete Fourier Transform (DFT), Odena et al. [48] discovered that adding insignificant zeros to a low-resolution image is equivalent to overlaying multiple spectra of the low-resolution image onto the high-frequency region of the resulting high-resolution image. This discrepancy causes the frequency spectrum of deepfake images to deviate from real images, making them distinguishable [13]. To extract forgery features in the frequency domain, we employ a two-dimensional Fast Fourier transform (2D FFT) on the input image A_i^x, resulting in the spectrum representation D_i. The backbone network generates a convolutional feature map $f_f \in \mathbb{R}^{H\times W\times C}$ using D_i as input. To direct the network's attention towards discriminative regions for classification, f_f is processed through an attention block, as illustrated in Fig. 4.

$$f_{att} = \phi(f_f) \tag{8}$$

$$F_{\mathrm{refined}f} = f_f \bigotimes f_{att} \tag{9}$$

The output of the attention block is element-wise multiplied with the f_f features, resulting in the refined feature map $F_{\mathrm{refined}f}$.

Once the frequency feature map $F_{\mathrm{refined}f}$ is obtained, these features are then combined with the spatial features F_{d1} from the first decoder layer of the U-Net. F_{d1} contains manipulation-aware features compared to the spatial features from the encoder's last layer. We utilize Bilinear Pooling (*BP*) to capture the comprehensive representation of these features. Bilinear Pooling merges features of different

dimensions and offers improved expressiveness compared to concatenation or element-wise product-based methods. Bilinear Pooling is computationally efficient and competitive with the best feature fusion strategies [49]. In the *BP* block, features from F_{d1} and $F_{\text{refined}f}$ are fused to compute the class probability.

3.3 Loss function

Four different loss functions, L1, L2, dice loss, and focal loss have been evaluated for manipulation localization network. L1 and L2 losses are commonly used in regression tasks, while dice loss and focal loss are typically utilized in classical segmentation tasks. Our findings show that L2 and L1 losses outperformed the segmentation losses, suggesting that the regression losses are more suitable for localized maps. Table 4 compares results using different loss functions. In addition, we combined U-Net with the classification network for training and employed binary cross-entropy loss. The overall loss is the weighted sum of the two activation losses, i.e., localization and classification loss:

$$L_{\text{comb}} = \rho_{\text{class}} L_{\text{class}} + \rho_{\text{localize}} L_{\text{localize}}. \tag{10}$$

The two weights ($\rho_{\text{class}}, \rho_{\text{localize}}$) are set to 1. This is because classification and localization tasks are equally important.

4 Experimental setup

4.1 Implementation details and evaluation settings

For all real/fake video frames, we employ MTCNN [50] to detect and crop the face region, saving the aligned facial images as inputs with a size of 224×224. ResNet [51] is used as a backbone network to extract spatial and frequency features. The model is trained using Adam optimizer [52] with an initial learning rate (LR) of $1e^{-4}$, $\beta 1 = 0.9$, $\beta 2 = 0.999$ $epsilon = 1e^{-08}$, and epsilon $= 1e^{-08}$. After 30 epochs, if the network does not improve, the learning rate drops to LR $\times$ 0.1. We train our models on NVIDIA GeForce RTX-3060 Ti GPUs with batch size 16. We used various data augmentation techniques to prevent overfitting and encourage the model to learn identity-independent features rather than solely focusing on face recognition. These techniques include flipping, rotating, contrast change, adding Gaussian noise, and compression to simulate diverse scenarios. In order to evaluate the efficacy of our suggested method, we utilize well-established metrics for detecting deepfakes. These metrics include Accuracy (Acc), which is used for assessing the performance of the model within the FaceForensics++ [28] dataset, as demonstrated in studies by [12, 51, 53–55]. Additionally, we employ area under the ROC curve (AUC) for evaluating

Table 1 Summary of face and expression swap datasets

Manipulation	Method	Dataset
Face swap	Deepfake and FaceSwap	CelebDF, DFDC-P, FF++ (DF, FS), DFD
Expression swap	Face2Face and Neural-Textures	FF++ (F2F, NT)

the model's performance across CelebDF [8], DFD [56], and DFDC-P [29] datasets, as shown in previous research by [10, 12, 16, 17, 28, 33, 37, 41, 57]. Finally, we also use Mean Intersection over Union (mIoU) for further evaluation. To ensure fair comparisons with other techniques, we calculate average metric scores for all frames within a video.

4.2 Datasets

We evaluated our proposed method on four benchmark datasets: FaceForensics++ [28], Celeb-DF [8], DFD [56], and DFDC-P [29], as summarized in Table 1.

1 **FaceForensics++ (FF++)** [28]: The dataset consists of 1000 original YouTube videos and 4000 fake videos generated using four manipulation algorithms: Deepfake (DF), FaceSwap (FS), Neural Textures (NT), and Face-to-Face (F2F). To ensure a balanced representation of real and fake data, 30 frames were selected from each fake video and 120 frames from each original video. Two different qualities of the dataset were used for training and testing: high quality (HQ) with a moderate compression ratio of 23 (C-23) and low quality (LQ) with a higher compression ratio of 40 (C-40). Higher compression results in lower video quality. The FF++ dataset now includes FaceShifter (FSH) face swapping videos, consisting of 10,000 fake videos created by manipulating real videos from the FF++ real videos.
2 **Deepfake Detection Challenge-Preview (DFDC-P)** [29]: The dataset includes 4113 face swap deepfake videos alongside 1131 original footages.
3 **Celeb-DF** [8]: This dataset contains 590 original videos and 5,639 fake face swap videos.
4 **Deepfake Detection Dataset (DFD)** [56]: Google and Jigsaw contributed to the dataset, which includes 363 real videos and over 3600 face swapped deepfake videos.

5 Results

5.1 Forgery detection results

We performed both inner and cross-data evaluations for the proposed approach. The training and testing sets were sourced from the same dataset for inner-dataset evaluation. In contrast, the cross-dataset evaluation involved training and testing on different datasets.

5.2 Inner-dataset evaluation

This section presents a comparison between our methods and established state-of-the-art techniques using an inner-dataset evaluation. Extensive research has been conducted on the task of deepfake detection [58]. Only methods trained on FF++ HQ (C-23) and tested on FF++ LQ (C-40) were considered for this comparison. Frame-level results on the FF++ dataset are reported for fair comparisons. Table 2 summarizes the accuracy results of different state-of-the-art detectors. The reported results for [12, 51, 53–55] are directly cited from [12] and [28]. Our proposed method achieves comparable or superior performance compared to the current

state-of-the-art approaches for low compression. Specifically, our approach demonstrates improved performance on Deepfake (DF) and Neural-Texture (NT) manipulations in both high-quality (C-23) and low-quality (C-40) videos. Our proposed solution shows slightly lower accuracy for F2F and FS manipulations. Despite the existence of ADD [43], and Multi-Task [41] approaches for localizing manipulation regions, our method outperformed them.

This shows that improved results are possible even for compressed video by combining frequency and spatial domain information, suggesting that frequency spectrum features are resilient to compression. Highly compressed videos often exhibit poor quality, leading to the weakening of several frequency components. The performance enhancement from the attention block at both the spatial and frequency backbones helped to prioritize features with higher classification importance. We evaluated the trained model's performance on the FaceShifter dataset and obtained an accuracy of 95.88% for C-23 and 89.88% for C-40 compressed videos. Figure 5 shows the ROC results for inner dataset evaluation on the FF++ dataset.

5.3 Cross-dataset evaluation

Generalization ability is a key indicator of algorithm superiority, often evaluated through a cross-dataset evaluation. However, it is more practical to evaluate across datasets because it is often difficult to determine which modification approach was used for the test data.

This section focuses on the framework's adaptability to unseen datasets during training, highlighting its transferability through cross-dataset evaluation. To evaluate the proposed method's transferability and enable fair comparisons, we trained it on FF++ with multiple manipulations and conducted tests on CelebDF [8] and DFDC-P [29]. Table 3 provides a comparison of AUC values with state-of-the-art face forgery detection methods.

Our method outperformed the most recent approaches in terms of AUC on the DFDC-P dataset and also performed well on Celeb-DF. In conclusion, CNN-based approaches [10, 12, 16, 17, 28, 33, 37, 41, 57] predominantly emphasize local features within facial images, lacking global information for comprehensive enhancement. As a result, these methods exhibit limited transferability when cross-evaluated on DFDC-P and Celeb-DF datasets. Moreover, CViT approach [39], which employs the convolutional

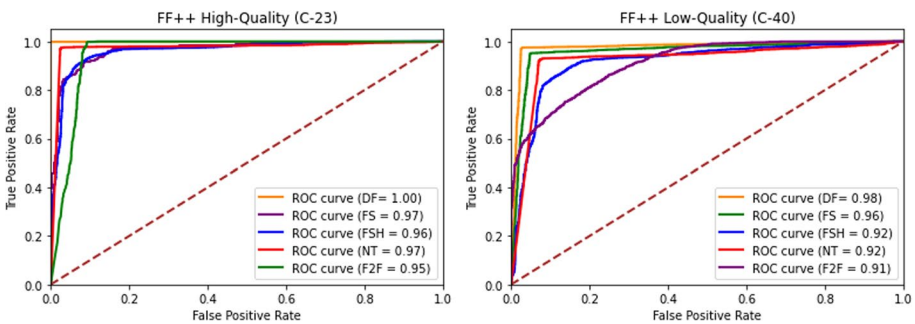

Fig. 5 ROC curves illustrating the classification performance on (HQ—C-23) and (LQ—C-40) compression qualities from the FF++ dataset

vision Transformer (ViT), experiences a decline in performance for inner-dataset evaluation on FF++ as compared to other state-of-the-art techniques [11, 16, 17, 24, 33, 37, 38, 43]. On the other hand, the recent state-of-the-art model MaDD [24] demonstrates relatively competitive performance in both within-dataset and cross-dataset evaluations compared to earlier approaches. In particular, our approach achieved a 4% higher AUC (area under the curve) compared to the highest reported approach, MesoNET [12], in evaluating the DFDC-P dataset. This improvement can be attributed to our method's emphasis on the input image's frequency and spatial components. While the two-branch [11] approach showcased superior transferability on the CelebDF dataset, our method outperformed it on DFDC-P dataset performance.

5.4 Manipulation localization results

The dual attention block scAB combines encoder and previous layer decoder features at each skip connection of ResU-Net. This integration allows the model to learn discriminative image features for manipulation localization while disregarding irrelevant pixels, as shown in Fig. 6. In training the proposed model, inverted FF++ ground truth masks are employed as input for fake and real faces. The inversion process involves representing the manipulated area with black pixels, while the real, unmanipulated area is depicted as white pixels as shown in Fig. 6. This inversion technique enhances the visualization of the model's predictions, as it distinctly highlights the regions where the face is manipulated.

As depicted in Fig. 1, facial expression modification usually occurs in regions such as the eyes, lips, and eyebrows. In deepfake face swap, all facial attributes except hair and ears are replaced. Figure 6 displays localized maps for original and fake images, demonstrating the model's effective learning of fake facial regions and accurate localization of potential manipulation pixels in each sample. The network maintains exceptional localization capabilities across all layers, particularly in accurately localizing the mouth, eyebrows, and eyes for expression changes. The network effectively localizes the facial region that has been transferred to the target image in face swap scenarios. In evaluating the face manipulation localization network, we examined four different loss functions using both original and fake images from the FF++ dataset for training and testing. The accuracy results, presented in Table 4, demonstrated that regression losses (L1 and L2) outperformed traditional segmentation losses in accurately localizing real and fake faces.

Next, we conduct a comparative analysis between our model and other approaches that utilize multi-task learning to enhance generalization capabilities. These approaches include LAE [62] and Multi-Task [41], and ADD [43]. These methods simultaneously perform forgery localization and classification. Following the same experimental setup as these methods, we train our model on the F2F (HQ) dataset and evaluate its effectiveness on both the F2F (HQ) and FS (HQ) datasets to measure its cross-dataset performance. The reported statistics for the competing methods can be found in the respective papers. As depicted in Table 5, our proposed method demonstrates better performance over the approaches [41, 43, 62] for cross-dataset evaluation.

To evaluate the impact of augmentation on analyzing unseen test sets, we trained our model on the FF++ (HQ) dataset comprising (DF, FSH, and real videos) with and

Table 2 Quantitative results in terms of ACC (%) on the FF++ [28] dataset were obtained for four different manipulation methods, including Deepfake (DF), Face2Face (F2F), FaceSwap (FS), and NeuralTextures (NT)

Method	Input	Mask	Face swap				Expression swap			
			DF (HQ)	DF (LQ)	FS (HQ)	FS (LQ)	NT (HQ)	NT (LQ)	F2F(HQ)	F2F (LQ)
Steg.Features+SVM [54]	RGB	N	77.12	65.58	79.51	68.93	76.94	60.69	74.68	60.58
Cozzolino et al. [51]	RGB	N	81.78	68.26	85.69	73.79	80.60	62.42	85.32	62.08
Bayar and Stamm [59]	RGB	N	90.18	80.95	93.14	82.52	86.04	72.38	94.93	76.83
MesoNet [12]	RGB	N	95.26	89.52	81.24	61.17	85.95	75.74	95.84	83.56
XceptionNet [28]	RGB	N	98.85	94.88	**98.23**	92.17	94.50	82.11	98.23	**91.56**
Multi-Task [41]	RGB	Y	93.92	85.77	–	–	88.05	80.67	92.77	82.31
Sun et al. [38]	RGB	N	–	69.1	–	68.1	–	60.8	–	65.7
SSTNet [14]	RGB	N	–	95.33	–	**94.09**	–	–	–	90.48
SPSL [53]	FREQ	N	–	93.48	–	92.26	–	76.78	-	86.02
ADD [43]	RGB	Y	97.45	–	97.20	–	90.84	–	**98.33**	–
Proposed method	RGB+FREQ	Y	**99.97**	**96.47**	97.88	93.88	**96.06**	**90.55**	95.97	90.92

This table summarizes the results, with "LQ" indicating low image quality, "HQ" indicating high image quality, "RGB" representing color images, and "FREQ" indicating frequency input. The best results are highlighted in bold font, while "–" indicates unavailable results

Table 3 Comparison of AUC (%) for cross-dataset evaluation on CelebDF [8] and DFDC-P [29], including results of other methods cited from [11, 24, 37, 38, 60, 61]

Method	FF++	CelebDF	DFDC-P
Two-stream [30]	70.10	53.80	61.4
MesoNET [12]	84.70	54.80	75.3
HeadPose [19]	47.3	54.6	55.9
VA-MLP [18]	66.4	55.0	61.9
FWA [57]	80.10	56.90	72.7
Xception-raw [28]	99.70	48.20	49.9
Xception-C23 [28]	99.70	65.30	72.2
Xception-C40 [28]	95.50	65.50	69.7
Capsule [10]	96.60	57.50	53.3
Multi-task [41]	76.30	54.30	53.6
Two-branch [11]	93.18	**73.41**	64.0
F^3Net [16]	98.10	65.17	70.1
EfficientNet [33]	99.70	64.29	70.12
Sun et al. [38]	99.3	64	69
GocNet [17]	97.55	67.43	–
ADD [43]	91.71	66.48	–
CViT [39]	91.08	63.60	67.3
MaDD [24]	**99.80**	67.44	67.1
FakePol [37]	94.7	61.2	72.5
Proposed method	97.78	68.25	**79.10**

Bold values indicate the best performace against the specfic dataset in each column

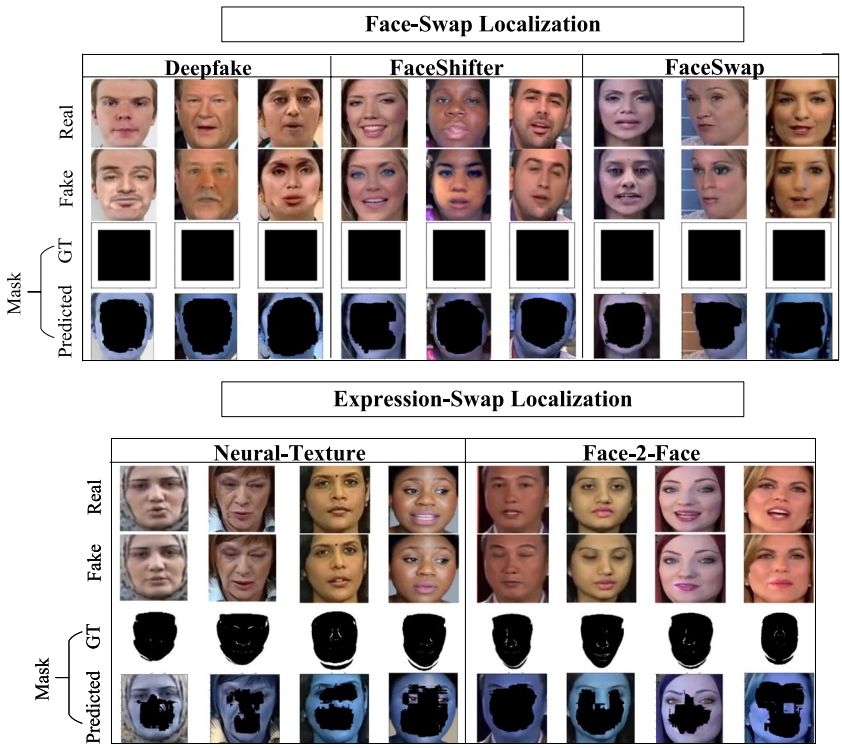

Fig. 6 First and second row from each deepfake manipulation type (DF, FS, FSH, NT, F2F) show the original images and manipulated ones, respectively. Third row shows the ground truth masks, while the bottom row represents the predicted mask from our proposed approach

Table 4 Comprehensive evaluation of localization loss functions on FF++ dataset

Loss-function	ACC (Fake) %	ACC (Real) %
Dice-loss	83.79	88.75
Focal-loss	87.78	89.30
L1	90.16	93.63
L2	93.77	95.45

Table 5 Facial manipulation localization performance in terms of accuracy on Face2Face and FaceSwap datasets with high video quality

Method	F2F (HQ)	FS (HQ)
Multi-task [41]	92.8	54.1
LAE [62]	90.9	63.2
ADD [43]	**98.33**	67.02
Proposed method	94.43	**70.04**

Bold values indicate the best performace against the specfic dataset in each column

Table 6 Localization network cross-data evaluation on DFDC-P with and without augmentation

Training approach	ACC (Fake-DFDC-P) %	ACC (Real-DFDC-P) %
FF + +(DF, FSH, Real)$_{w-aug}$	61.98	59.24
FF + +(DF, FSH, Real)$_{w/o-aug}$	53.73	52.98

without data augmentation. Subsequently, we assessed the model's performance on the DFDC-P dataset. To showcase the effectiveness comprehensively, we conducted tests on both the DFDC-P(real) and DFDC-P (fake) datasets, as presented in Table 6. The table showcases the effectiveness of augmentation in improving cross-data performance. The first row, labeled "w-aug", corresponds to the training approach with augmentation, while the second row, labeled "w/o-aug", refers to training without data augmentation.

Upon analyzing the evaluation results within the dataset in Table 2 and the cross-data assessment results in Table 3, noticeable performance variations are observed among unseen datasets. This substantiates the challenges posed by the distribution gap between the seen and unseen datasets regarding generalization accuracy. Our future study will focus on exploring additional features, such as background context or voice, to examine if they can contribute to further reducing the generalization gap. In particular, we will investigate the potential of incorporating a limited amount of data from unseen datasets for fine-tuning the model.

6 Ablation study

We independently performed ablation experiments for each localization and detection branch with different network configurations to assess the effectiveness of each component in the proposed approach.

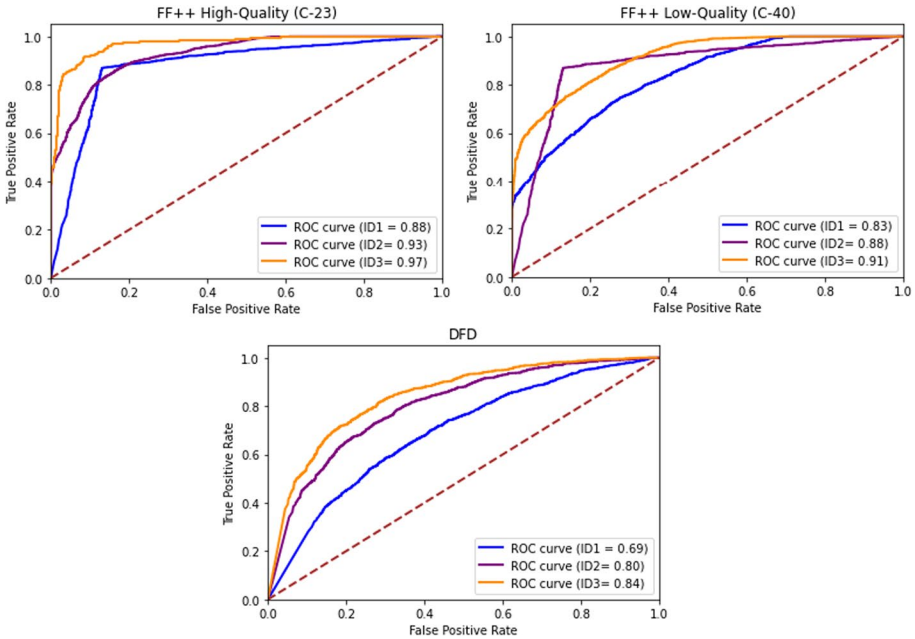

Fig. 7 The detection performance of different features on FF++ and DFD datasets. ID1 only uses frequency domain features, ID2 uses spatial and frequency domain features without a ScAB block, and ID3 includes a ScAB block

Table 7 Component analysis on the proposed detection branch for the high-quality (HQ), low-quality (LQ) FF++, and DFD datasets. Each component is gradually incorporated and evaluated to compare the ACC (%) results

ID	Frequency features	Spatial features	SCAB	FF++ (C-23)	FF++ (C-40)	DFD
1	✓			87.62	82.23	60.22
2	✓	✓		92.50	89.33	72.91
4	✓	✓	✓	97.33	92.50	76.49

6.1 Deepfake manipulation detection network

We quantitatively assessed the significance of the detection model's component in understanding the detection efficiency of the proposed network. We compared the output from (a) the detection branch trained with only frequency domain features, (b) the detection branch using features from the frequency and spatial domain with no ScAB block, and (c) a combination of both frequency and spatial domain with ScAB block. To show this, all models are trained on FF++ (HQ) and evaluated on FF++ (LQ) and DFD datasets.

In contrast to combining information from the spatial and frequency domains, we find that employing features from the frequency domain alone does not yield satisfactory results. One should not discard all spatial information and depend solely on frequency domain parameters for classification. Instead, combining both domains boosts performance considerably. A simple hard combination of features from both domains using a Bilinear Pooling layer enhances the performance. However, in this case, there is no

information about the manipulation location on the face, giving limited room for information flow between both domains.

We used input from the first decoder layer with ScAB block to show the impact of integrating pixel-wise forgery localized spatial information with frequency domain features. This permits only the altered spatial pixels to be shared with the detection branch rather than features from the entire face, thereby learning a better optimal combination of shared representations from both feature domains. Table 7 and Fig. 7 illustrate that the optimal results for within dataset and cross datasets are achieved through the combination of frequency and localized spatial domain features.

6.2 Deepfake manipulation localization network

In this ablation study, we aimed to investigate the impact of spatial and channel attention on the performance of the localization branch. We conducted quantitative experiments and provided visualization to demonstrate the importance of the ScAB block. We compared the performance of three models trained with different components:

- **Model-A**: Localization branch with spatial channel attention block (scAB).
- **Model-B**: Localization branch with spatial attention block only (sAB).
- **Model-C**: Localization branch without any attention block (W/O AB).

Localization results from Model-A, Model-B, and Model-C on FF++ dataset

We conducted an ablation study on the localization network using FF++ (C23) and FF++ (C40) for training and evaluation. The results are summarized in Table 8. The study revealed that including attention blocks significantly improved the performance of the localization branch. Model-C, with no attention block, exhibited the lowest performance. On the FF++ (C23) dataset, Model-B, relying on a spatial attention block (sAB) only, outperformed Model-A. However, it was observed that on highly compressed images from the FF++ (C40) dataset, where a substantial amount of information was lost due to high compression, the sAB was outperformed by ScAB. This outcome can be attributed to the compression affecting both local image features and their surroundings. The reliance of Model-A on local image features and Model-B solely on spatial attention to expand the areas of local image features can lead to inaccuracies in attention weights and mislocalizations, particularly due to the loss of crucial details in highly compressed images. In contrast, the scAB, capturing not only global but also contextual information through channel attention, proved to be especially effective for the challenging FF++ (C40) dataset. Consequently, Model-A achieved the highest performance on FF++

Table 8 Comprehensive evaluation of localization performance using AUC and mIoU metrics for all models on the FF++ dataset with two levels of video quality

Models	FF++ (C-23)		FF++ (C-40)	
	mIoU (%)	AUC (%)	mIoU (%)	AUC (%)
Model-A	85.67	95.06	82.21	93.62
Model-B	85.79	95.94	80.01	91.20
Model-C	80.42	93.25	74.31	90.57

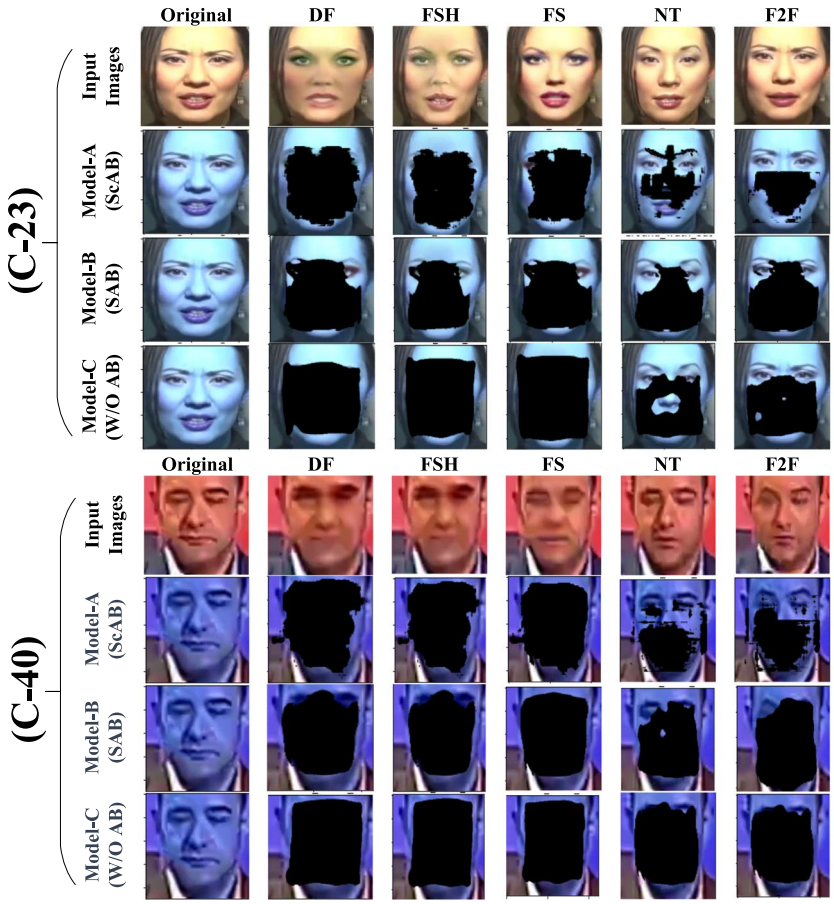

Fig. 8 Localization results from Model-A, Model-B, and Model-C on FF++ dataset

(C40). Figure 8 showcases the localization results of each model. Model-A, equipped with both spatial and channel attention blocks, exhibited more focused attention on the manipulated pixel regions, incorporating local, global, and contextual information through spatial and channel attention from the image. Furthermore, the localization results obtained with ScAB successfully highlight the forged pixels of the manipulated regions, even in low-quality FF++ faces. This highlights the effectiveness of ScAB in addressing the challenges posed by low-quality images.

7 Conclusion

This paper addresses the problem of detection and localization of faces in deepfake images using a multi-task learning approach. Our proposed method incorporates an attention mechanism to process the feature maps for both detection and localization tasks. By enabling information exchange between these tasks, we observed an overall improvement in the network's performance, particularly for unseen datasets. To enhance the performance and provide localization of face forgery, we introduce a strategy involving the combination of the encoder and preceding layer decoder with dual attention block scAB. This approach localizes the manipulated facial regions at the pixel level. Through extensive experiments on three deepfake benchmarks,

we demonstrate that our model tends to focus on the forgery regions instead of unwanted biases and artifacts, leading to more accurate predictions. Furthermore, we empirically show that the utilization of multiple attention blocks enhances the model's ability to localize manipulated regions. This improvement contributes to achieving state-of-the-art performance in forgery detection. With the improved visual quality of deepfake generated faces, the detection problem remains highly challenging, resulting in a generalization gap between training and unseen datasets created by other approaches. In our future research, we aim to address this generalization gap by combining transfer learning with additional strategies, creating a comprehensive framework to further narrow this disparity.

Acknowledgements
The authors would like to thank the Research Management Center of Universiti Teknologi Malaysia for managing the fund under vol. no. 4C396..

Funding
This research received no external funding.

Availability of data and materials
Celeb-DF, FF++, DFD and DFDC-P datasets are available upon request at [https://github.com/yuezunli/celeb-deepfakeforensics, accessed on March 2021] and [https://github.com/ondyari/FaceForensics, accessed on January 2021], respectively.

Declarations

Competing interests
The authors declare that they have no competing interests.

Received: 26 October 2022 Accepted: 8 August 2023
Published online: 18 August 2023

References

1. J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, M. Nießner, Face2face: Real-time face capture and reenactment of rgb videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2387–2395 (2016)
2. Y. Lu, Y.-W. Tai, C.-K. Tang, Attribute-guided face generation using conditional cyclegan. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 282–297 (2018)
3. H. Kim, P. Garrido, A. Tewari, W. Xu, J. Thies, M. Niessner, P. Pérez, C. Richardt, M. Zollhöfer, C. Theobalt, Deep video portraits. ACM Trans. Graph. (2018). https://doi.org/10.1145/3197517.3201283
4. L. Li, J. Bao, H. Yang, D. Chen, F. Wen, FaceShifter: towards high fidelity and occlusion aware face swapping (2020). arXiv:1912.13457
5. S. Lu, FaceSwap-GAN. https://github.com/shaoanlu/faceswap-GAN. Accessed: 2022-01-30
6. C. Gosse, J. Burkell, Politics and porn: how news media characterizes problems presented by deepfakes. Crit. Stud. Media Commun. **37**(5), 497–511 (2020). https://doi.org/10.1080/15295036.2020.1832697
7. M. Westerlund, The emergence of deepfake technology: a review. Technol. Innovat. Manag. Rev. **9**(11) (2019)
8. Y. Li, X. Yang, P. Sun, H. Qi, S. Lyu, Celeb-df: a large-scale challenging dataset for deepfake forensics. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3204–3213 (2020). https://doi.org/10.1109/CVPR42600.2020.00327
9. J. Yang, A. Li, S. Xiao, W. Lu, X. Gao, Mtd-net: learning to detect deepfakes images by multi-scale texture difference. IEEE Trans. Inf. Forensics Secur. **16**, 4234–4245 (2021). https://doi.org/10.1109/TIFS.2021.3102487
10. H.H. Nguyen, J. Yamagishi, I. Echizen, Capsule-forensics: Using capsule networks to detect forged images and videos. In: ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2307–2311 (2019). https://doi.org/10.1109/ICASSP.2019.8682602
11. I. Masi, A. Killekar, R.M. Mascarenhas, S.P. Gurudatt, W. AbdAlmageed, Two-branch recurrent network for isolating deepfakes in videos, in *Computer Vision—ECCV 2020*. ed. by A. Vedaldi, H. Bischof, T. Brox, J.-M. Frahm (Springer, Cham, 2020), pp.667–684
12. D. Afchar, V. Nozick, J. Yamagishi, I. Echizen, Mesonet: a compact facial video forgery detection network. In: 2018 IEEE International Workshop on Information Forensics and Security (WIFS), pp. 1–7 (2018). https://doi.org/10.1109/WIFS.2018.8630761
13. R. Durall, M. Keuper, J. Keuper, Watch your up-convolution: Cnn based generative deep neural networks are failing to reproduce spectral distributions. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7887–7896 (2020). https://doi.org/10.1109/CVPR42600.2020.00791

14. X. Wu, Z. Xie, Y. Gao, Y. Xiao, Sstnet: detecting manipulated faces through spatial, steganalysis and temporal features. In: ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2952–2956 (2020). https://doi.org/10.1109/ICASSP40776.2020.9053969
15. D. Güera, E.J. Delp, Deepfake video detection using recurrent neural networks. In: 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 1–6 (2018). https://doi.org/10.1109/AVSS.2018.8639163
16. Y. Qian, G. Yin, L. Sheng, Z. Chen, J. Shao, Thinking in frequency: face forgery detection by mining frequency-aware clues, in *Computer Vision—ECCV 2020*. ed. by A. Vedaldi, H. Bischof, T. Brox, J.-M. Frahm (Springer, Cham, 2020), pp.86–103
17. Z. Guo, G. Yang, D. Zhang, M. Xia, Rethinking gradient operator for exposing ai-enabled face forgeries. Expert Syst. Appl. **215**, 119361 (2023). https://doi.org/10.1016/j.eswa.2022.119361
18. F. Matern, C. Riess, M. Stamminger, Exploiting visual artifacts to expose deepfakes and face manipulations. In: 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW), pp. 83–92 (2019). https://doi.org/10.1109/WACVW.2019.00020
19. X. Yang, Y. Li, S. Lyu, Exposing deep fakes using inconsistent head poses. In: ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 8261–8265 (2019). https://doi.org/10.1109/ICASSP.2019.8683164
20. B. Xu, J. Liu, J. Liang, W. Lu, Y. Zhang, Deepfake videos detection based on texture features. Comput. Mater. Continua **68**(1), (2021)
21. F. Lugstein, S. Baier, G. Bachinger, A. Uhl, Prnu-based deepfake detection. In: Proceedings of the 2021 ACM Workshop on Information Hiding and Multimedia Security, pp. 7–12 (2021)
22. Y. Zhu, Q. Li, J. Wang, C. Xu, Z. Sun, One shot face swapping on megapixels. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4832–4842 (2021). https://doi.org/10.1109/CVPR46437.2021.00480
23. A. Groshev, A. Maltseva, D. Chesakov, A. Kuznetsov, D. Dimitrov, Ghost-a new face swap approach for image and video domains. IEEE Access **10**, 83452–83462 (2022). https://doi.org/10.1109/ACCESS.2022.3196668
24. H. Zhao, T. Wei, W. Zhou, W. Zhang, D. Chen, N. Yu, Multi-attentional deepfake detection. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2185–2194 (2021). https://doi.org/10.1109/CVPR46437.2021.00222
25. J. Li, H. Xie, L. Yu, X. Gao, Y. Zhang, Discriminative feature mining based on frequency information and metric learning for face forgery detection. IEEE Transactions on Knowledge and Data Engineering, 1–1 (2021). https://doi.org/10.1109/TKDE.2021.3117003
26. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks. Commun. ACM **63**(11), 139–144 (2020). https://doi.org/10.1145/3422622
27. J. Frank, T. Eisenhofer, L. Schönherr, A. Fischer, D. Kolossa, T. Holz, Leveraging frequency analysis for deep fake image recognition. In: International Conference on Machine Learning, pp. 3247–3258 (2020). PMLR
28. A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, M. Nießner, Faceforensics++: learning to detect manipulated facial images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1–11 (2019)
29. B. Dolhansky, R. Howes, B. Pflaum, N. Baram, C.C. Ferrer, The deepfake detection challenge (dfdc) preview dataset. arXiv preprint arXiv:1910.08854 (2019)
30. P. Zhou, X. Han, V.I. Morariu, L.S. Davis, Two-stream neural networks for tampered face detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1831–1839 (2017). https://doi.org/10.1109/CVPRW.2017.229
31. T. Jung, S. Kim, K. Kim, Deepvision: Deepfakes detection using human eye blinking pattern. IEEE Access **8**, 83144–83154 (2020). https://doi.org/10.1109/ACCESS.2020.2988660
32. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255 (2009). https://doi.org/10.1109/CVPR.2009.5206848
33. M. Tan, Q. Le, EfficientNet: Rethinking model scaling for convolutional neural networks. In: Chaudhuri, K., Salakhutdinov, R. (eds.) Proceedings of the 36th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 97, pp. 6105–6114. PMLR, (2019). https://proceedings.mlr.press/v97/tan19a.html
34. S. Waseem, S.R. Abu-Bakar, Z. Omar, B.A. Ahmed, S. Baloch, A multi-color spatio-temporal approach for detecting deepfake. In: 2022 12th International Conference on Pattern Recognition Systems (ICPRS), pp. 1–5 (2022). https://doi.org/10.1109/ICPRS54038.2022.9853853
35. E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, P. Natarajan, Recurrent convolutional strategies for face manipulation detection in videos. Interfaces (GUI) **3**(1), 80–87 (2019)
36. R.D. Lopez, M. Keuper, F.-J. Pfreundt, J. Keuper, Unmasking DeepFakes with simple Features (2019)
37. L. Tian, H. Yao, M. Li, Fakepoi: A large-scale fake person of interest video detection benchmark and a strong baseline. IEEE Transactions on Circuits and Systems for Video Technology, 1–1 (2023). https://doi.org/10.1109/TCSVT.2023.3269742
38. K. Sun, H. Liu, Q. Ye, Y. Gao, J. Liu, L. Shao, R. Ji, Domain general face forgery detection by learning to weight. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 2638–2646 (2021)
39. D. Wodajo, S. Atnafu, Deepfake Video Detection Using Convolutional Vision Transformer (2021). arXiv:2102.11126
40. Y. Hua, R. Shi, P. Wang, S. Ge, Learning patch-channel correspondence for interpretable face forgery detection. IEEE Trans. Image Process. **32**, 1668–1680 (2023). https://doi.org/10.1109/TIP.2023.3246793
41. H.H. Nguyen, F. Fang, J. Yamagishi, I. Echizen, Multi-task learning for detecting and segmenting manipulated facial images and videos. In: 2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS), pp. 1–8 (2019). https://doi.org/10.1109/BTAS46853.2019.9185974
42. L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, B. Guo, Face x-ray for more general face forgery detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5001–5010 (2020)
43. P. Liu, Y. Lin, Y. He, Y. Wei, L. Zhen, J.T. Zhou, R.S.M. Goh, J. Liu, Automated deepfake detection. arXiv preprint arXiv:2106.10705 (2021)

44. H. Dang, F. Liu, J. Stehouwer, X. Liu, A.K. Jain, On the detection of digital face manipulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5781–5790 (2020)
45. B. Niu, W. Wen, W. Ren, X. Zhang, L. Yang, S. Wang, K. Zhang, X. Cao, H. Shen, Single image super-resolution via a holistic attention network, in *Computer Vision—ECCV 2020*. ed. by A. Vedaldi, H. Bischof, T. Brox, J.-M. Frahm (Springer, Cham, 2020), pp.191–207
46. S. Woo, J. Park, J.-Y. Lee, I.S. Kweon, Cbam: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–19 (2018)
47. T. Park, J.-Y. Zhu, O. Wang, J. Lu, E. Shechtman, A. Efros, R. Zhang, Swapping autoencoder for deep image manipulation. In: H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, H. Lin, (eds.) Advances in Neural Information Processing Systems, vol. 33, pp. 7198–7211. Curran Associates, Inc., (2020). https://proceedings.neurips.cc/paper/2020/file/50905d7b2216bfeccb5b41016357176b-Paper.pdf
48. A. Odena, V. Dumoulin, C. Olah, Deconvolution and checkerboard artifacts. Distill **1**(10), 3 (2016)
49. T.-Y. Lin, A. RoyChowdhury, S. Maji, Bilinear cnn models for fine-grained visual recognition. In: 2015 IEEE International Conference on Computer Vision (ICCV), pp. 1449–1457 (2015). https://doi.org/10.1109/ICCV.2015.170
50. K. Zhang, Z. Zhang, Z. Li, Y. Qiao, Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process. Lett. **23**(10), 1499–1503 (2016). https://doi.org/10.1109/LSP.2016.2603342
51. D. Cozzolino, G. Poggi, L. Verdoliva, Recasting residual-based local descriptors as convolutional neural networks: an application to image forgery detection. In: Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security, pp. 159–164. Association for Computing Machinery, New York, NY, USA (2017). https://doi.org/10.1145/3082031.3083247
52. D.P. Kingma, J. Ba, Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
53. H. Liu, X. Li, W. Zhou, Y. Chen, Y. He, H. Xue, W. Zhang, N. Yu, Spatial-phase shallow learning: Rethinking face forgery detection in frequency domain. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 772–781 (2021). https://doi.org/10.1109/CVPR46437.2021.00083
54. J. Fridrich, J. Kodovsky, Rich models for steganalysis of digital images. IEEE Trans. Inf. Forensics Secur. **7**(3), 868–882 (2012). https://doi.org/10.1109/TIFS.2012.2190402
55. J.-Y. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 2242–2251 (2017). https://doi.org/10.1109/ICCV.2017.244
56. D.S. Nigeria, DeepFake-Detection Dataset. https://github.com/DataScienceNigeria/Fake-Detection-dataset-for-deepfake-from-Google-and-Jigsaw. Accessed: 2022-04-30
57. Y. Li, S. Lyu, Exposing deepfake videos by detecting face warping artifacts. arXiv preprint arXiv:1811.00656 (2018)
58. R. Tolosana, S. Romero-Tapiador, R. Vera-Rodriguez, E. Gonzalez-Sosa, J. Fierrez, Deepfakes detection across generations: analysis of facial regions, fusion, and performance evaluation. Eng. Appl. Artif. Intell. **110**, 104673 (2022). https://doi.org/10.1016/j.engappai.2022.104673
59. B. Bayar, M.C. Stamm, A deep learning approach to universal image manipulation detection using a new convolutional layer, pp. 5–10. Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/2909827.2930786
60. S.S. Khalil, S.M. Youssef, S.N. Saleh, icaps-dfake: an integrated capsule-based model for deepfake image and video detection. Future Internet (2021). https://doi.org/10.3390/fi13040093
61. C. Fosco, E. Josephs, A. Andonian, A. Lee, X. Wang, A. Oliva, Deepfake caricatures: amplifying attention to artifacts increases deepfake detection by humans and machines. arXiv preprint arXiv:2206.00535 (2022)
62. M. Du, S. Pentyala, Y. Li, X. Hu, Towards generalizable deepfake detection with locality-aware autoencoder. In: Proceedings of the 29th ACM International Conference on Information and Knowledge Management. CIKM '20, pp. 325–334. Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3340531.3411892

Publisher's Note