Expert Systems With Applications 210 (2022) 118423

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

ViXNet: Vision Transformer with Xception Network for deepfakes based video and image forgery detection

Shreyan Ganguly [a], Aditya Ganguly [b], Sk Mohiuddin [c,*], Samir Malakar [c], Ram Sarkar [b]

[a] Department of Construction Engineering, Jadavpur University, Kolkata, India
[b] Department of Computer Science and Engineering, Jadavpur University, Kolkata, India
[c] Department of Computer Science, Asutosh College, Kolkata, India

ARTICLE INFO

Keywords:
Deepfakes
FaceSwap
Soft attention
Vision transformer
Forgery detection
Xception model

ABSTRACT

With the advent of image generative technologies, there is a huge growth in the development of facial manipulation techniques that allow people to easily modify media data like videos and images by changing the identity or facial expression of the target person with another person's face. Colloquially, these manipulated videos and images are termed "deepfakes". As a result, every piece of content in digital media comes with a question — is this authentic? Hence, there is an unprecedented need for a competent deepfakes detection method. The rapid changes in forging methods make this a very challenging task and thus generalization of the detection methods is also of utmost required. However, the generalization strengths of the prevailing deepfakes detection methods are not satisfactory. In other words, these models perform well when trained and tested on the same dataset but fail to perform satisfactorily when models are trained on one dataset and tested on another. The most modern deep learning aided deepfakes detection techniques looked for a consistent pattern among the leftover artifacts in specific facial regions of the target face rather than the entire face. To this end, we propose a Vision Transformer with Xception Network (ViXNet) to learn the consistency of these almost imperceptible artifacts left by deepfaking methods on the entire facial region. The ViXNet comprises two branches — one tries to learn inconsistencies among local face region specifics by combining patch-wise self-attention module and vision transformer, and the other generates global spatial features using a deep convolutional neural network. To assess the performance of ViXNet, we evaluate it using two different experimental setups — intra-dataset and inter-dataset when using three standard deepfakes video datasets, namely FaceForensics++, and Celeb-DF (V2) and one deepfakes image dataset called Deepfakes. We have attained 98.57% (83.60%), 99.26% (74.78%), and 98.93% (75.13%) AUC scores using intra(inter)-dataset experimental setups on FaceForensics++, Celeb-DF (V2), and Deepfakes datasets respectively. Additionally, we have evaluated ViXNet on the Deepfake Detection Challenge (DFDC) dataset and we have obtained 86.32% AUC score and 79.06% F1-score on the said dataset. Performances of the proposed model are comparable to state-of-the-art methods. Besides, the obtained results ensure the robustness and the generalization ability of the proposed model.

1. Introduction

Images have been perceived as an authentic and reliable method of information transfer for a long time. However, with the emergence of digital technologies, the availability of a huge amount of data, and the continued increase in computing capability and memory capacity, it has become very easy to train even very large deep learning models for various image processing and pattern recognition tasks. This has also made it very easy to manipulate images and videos using some sophisticated deep learning models, mostly by using generative adversarial networks (GANs). As a side effect of this, several algorithms have been developed over the past few years to generate increasingly realistic media data having manipulated faces, conversationally called deepfakes. Formally speaking, deepfakes are synthetic media data like videos and images, where the malicious users replace a target person's face with someone else. There are two main categories of deepfakes techniques: (i) face swapping and (ii) face reenactment. In the face swap methods, the face of one individual is cropped and morphed to fit it in the context of another individual as depicted in Fig. 1, while

* Corresponding author.
E-mail addresses: gshreyan16@gmail.com (S. Ganguly), aditya.ganguly3145@gmail.com (A. Ganguly), myselfmohiuddin@gmail.com (S. Mohiuddin), malakarsamir@gmail.com (S. Malakar), ramjucse@gmail.com (R. Sarkar).

https://doi.org/10.1016/j.eswa.2022.118423
Received 27 January 2022; Received in revised form 5 July 2022; Accepted 3 August 2022
Available online 8 August 2022

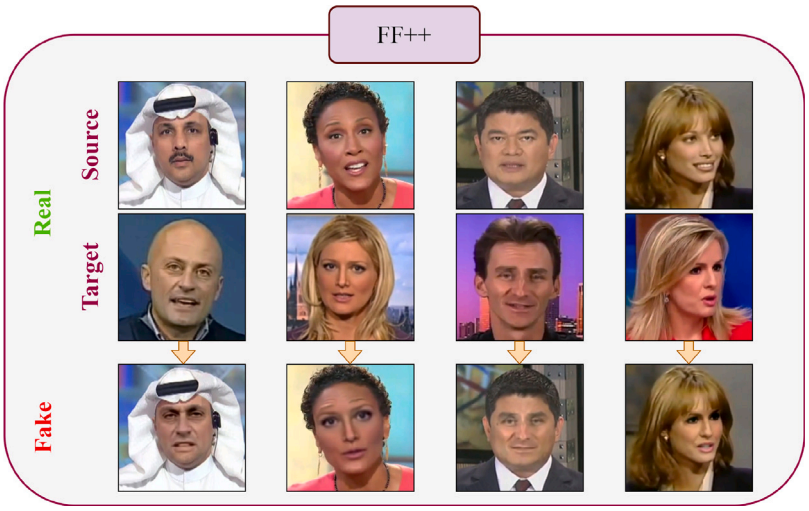


Fig. 1. Some examples of deepfakes cropped face images taken from FaceForensics++ (FF++) dataset (Rossler et al., 2019).

in face reenactment methods, the facial expression of an individual is changed. Fig. 1 depicts some sample deepfakes images (last row) that are generated by morphing the source (top row) and target (middle row) images. These newly generated manipulated face images are very close to real images and hard to say whether they are real or not. As a result, such generated deepfakes may be easily used for nefarious purposes such as spreading fake news, financial fraud, and hoaxes. Therefore, it is quite apparent that there is an utmost need for robust and effective deepfakes detection algorithms.

Recently convolutional neural network (CNN) models, in particular deep CNN models, have been successfully applied to the field of computer vision. Given their translation invariant property, they have been very successful in capturing useful features from images. Leveraging this feature extracting power of deep CNNs, several deep learning models have been proposed in the past for the detection of deepfakes. For example, Afchar et al. (2018) proposed the MesoNet architecture for detecting macroscopic artifacts that are present in manipulated videos. Nirkin et al. (2021) designed a hybrid architecture where the individual faces and their surrounding context are processed separately, and finally fused both to make the prediction. Videos have also been treated as time-series data to capture the inter-connectivity among different frames as proposed by Chintha et al. (2020). It is worth mentioning that the Xception network introduced by Chollet (2017), which uses depth separable convolution layers, has not only been used very successfully as a baseline in various computer vision tasks but also provided state-of-the-art deepfakes detection performances on many publicly available deepfakes datasets when a single CNN model is considered (Li et al., 2020).

While deep CNN models can very effectively detect local artifacts, modern deepfakes generation techniques can create a wide range of artifacts, from ones that are local to those that span the entire image. Moreover, there is a large diversity in the types of artifacts produced owing to the different types of generation techniques available. The existing deep learning based methods sometimes fail to deal with such diversity as found in Li et al. (2020). To encounter such diversity in the faking process, many researchers as in Yu et al. (2022) incorporated differently generated face manipulated videos into their evaluation process. However, it may not be a viable solution considering the large spectrum faking methods. Hence, there is a pressing need to design methods that can effectively deal with a large variety of fake artifacts.

Following the seminal paper by Dosovitskiy et al. (2021) which successfully applied transformers, formerly used in the Natural Language Processing (NLP) domain, to the task of image classification, there has been an explosion in their use in the field of computer vision. The Vision Transformer (known as ViT Dosovitskiy et al., 2021) has been successfully applied to tasks such as image classification (Dosovitskiy et al., 2021), image segmentation (Chen, Lu et al., 2021) and object detection (Carion et al., 2020). This success is due to the ability of transformers to capture global correlation among different image regions. In most common practice, unaltered patches of fixed dimensions are used as image regions. This approach helps to find the consistencies of various image regions. However, this solution can be optimized by altering the image regions fed into the transformer, which serves as pre-processing to suppress image content and highlight manipulation traces.

Considering the above mentioned facts, in this work, we propose a deep learning model, which uses patch-level learnable masking to convert local image regions into masked features. These masked features are further fed into the ViT model to capture global inconsistencies among the generated local masked features. We also consider global image features, by using a deep CNN model, called Xception. Next, the differently generated features are combined to obtain the final features which express global inconsistencies of local features, and image spatial features. These generated features become very useful in the deepfakes detection domain owing to their ability to learn the leftover artifacts, either local or global, in deepfakes videos/images effectively. We call this model hereafter as Vision Transformer with Xception Network or in short ViXNet. In addition to that the proposed model can be used to detect face forgery automatically by integrating with some user interfaces like Khan et al. (2022) and Yousaf et al. (2019). We evaluate the performance of the proposed ViXNet model on three standard datasets — FaceForensics++ (Rossler et al., 2019), Celeb-DF (Li et al., 2020), and Deepfakes Image Dataset (Afchar et al., 2018) to ensure the robustness and generalizability of the same. We perform experiments in intra-dataset and inter-dataset scenarios, to confirm whether the model can learn artifacts generated by a specific forging algorithm as well as by some unknown forging algorithms.

The highlights of present work can be summarized as follows:

- We propose a deep learning based deepfakes detection model which exploits leftover artifacts efficiently.
- A patch-wise self-attention module has been used with a vision transformer or ViT to learn image artifacts consistently present in different facial regions.
- Global image features have been fused with the local ones to generate a hybrid feature representation expressing global inconsistencies of local masked features and global image features.
- An extensive evaluation on some standard datasets confirms its superiority in terms of robustness and generalizability over many state-of-the-art methods.

The rest of the paper is organized in the following manner. In Section 2, an overview of deep learning methods used for deepfakes detection in the past has been presented. Section 3 states the motivation behind the present work along with the contributions made in this work in brief. Section 4 describes the working principle of the proposed model. In Section 5, we evaluate our method in extracting global inconsistencies, furthermore, the related datasets and experimental protocol have been outlined. Lastly in Section 6, we have provided some conclusive remarks about our work.

2. Related work

With the arrival of extremely powerful GAN models (Choi et al., 2018; Karras et al., 2019), detection of deepfakes becomes a challenging task. The earlier works for face manipulation detection relied mostly on different hand-crafted features that used different image properties like Discrete Fourier Transform (DFT) (Durall et al., 2019), facial landmarks (Yang et al., 2019), the camera captured noises (Koopman et al., 2018) to extract the patterns of forged artifacts. These methods mostly tried to expose the artifacts left by the face manipulation methods. However, with the advancement in GAN models and different sophisticated multimedia data compression techniques presence of these forging artifacts becomes almost invisible. Thus, detecting the leftover artifacts became challenging and the feature engineering based methods failed to provide satisfactory outcomes. On the other hand, evolution in CNN models makes the feature extraction process much easier due to their auto learnable feature extraction capability. These models produce a better performance compared to the hand-crafted feature based methods. As a result, a switch from hand-crafted feature based methods to deep feature aided methods has been observed for this purpose.

Several researchers (Afchar et al., 2018; Bayar & Stamm, 2018; Guo et al., 2021; Li et al., 2020; Rossler et al., 2019), in the era of deep features aided models, mostly relied on single CNN based model for learning artifacts present in deepfakes videos/images. Some of them (Afchar et al., 2018; Bayar & Stamm, 2018; Guo et al., 2021) trained their models from scratch (i.e., without using any pre-trained weights) during image manipulation detection. For example, Afchar et al. (2018) proposed two CNN architectures, namely (1) MesoNet-4, having 4 convolutional blocks followed by a fully connected layer, and (2) MesoInception-4, inspired by the Inception module (Szegedy et al., 2015), where the first two convolutional blocks of MesoNet-4 were replaced by Inception model's first two convolutional blocks. Guo et al. (2021) attempted to learn subtle manipulation traces by highlighting manipulation content and suppressing image content. Using convolutional layers, an intermediate feature map is generated and the original image is subtracted from this map. This effectively filters out unnecessary parts of the image and allows the remaining network to focus on the important details. In another work, Bayar and Stamm (2018) designed a CNN aided general purpose forensic technique, where a constrained convolutional layer was introduced to suppress the content of the image and learn manipulation detection features subsequently. These models performed well but the primary limitation of such methods is that they required a large number of samples to properly train the models to obtain satisfactory results. Additionally, these methods require a resource-heavy system to train from the scratch.

The widely accepted transfer learning concept curbs the need for voluminous data to train a CNN model and thus reduces the need for resource-heavy systems. The use of this transfer learning concept observed great success in many image classification problems and consequently, researchers started utilizing this concept to design CNN based deepfakes detection techniques. For example, Kumar et al. (2020) partitioned the cropped face into 4 non-overlapping blocks first and then fed them into the parallel ResNet-18 network. In the end, they combined the deep features from the four CNNs into a single feature vector to recognize the fake image. The objective of the work was to use local features for deepfakes detection, however, it did not investigate the relation among the features extracted from these local regions. In another work, Nguyen et al. (2019) used a part of a pre-trained VGG-19 network followed by a capsule network which is a parallel CNN network to distinguish manipulated faces. Güera and Delp (2018) made an attempt where they first extracted frame level features using a self-made CNN architecture and then fed these features into a recurrent neural network (RNN) for manipulated face detection. Unlike the other methods which resort to a single video frame, Amerini et al. (2019) used optical flow among the video frames to identify possible inter-frame abnormality which is then used in a VGG-16 based CNN model to identify forged videos. These methods condensed the need for voluminous training data and provided good results. However, these detectors tend to search artifacts left by some specific forging technique on a limited face region and thus they fail to learn the complete characteristics of these artifacts over the entire facial region (Wang & Deng, 2021). It is to be noted that the nature of the left-over artifacts varies from one deepfaking technique to another. As a result, these detectors failed to perform satisfactorily on deepfakes that are generated by some unknown technique. This fact had been proven by Li et al. (2020) who evaluated 8 existing CNN aided methods for which code and pre-trained weights were available publicly on their prepared dataset, known as Celeb-DF. In their experiments, they showed the Xception model has better generalization ability than other existing models.

To counter this issue Nirkin et al. (2021) designed a complex model for identity manipulation detection where confidence scores from a face recognizer were used in the detection model. In this work, the authors first partitioned the cropped faces into a tight facial region and context part (leaving out the face part) and then passed these two parts through the face recognition model to obtain face recognition scores. These scores were then stacked with the scores from the Xception (Chollet, 2017) based deepfakes detection model, where the entire cropped face was used. Finally, these combined scores were passed through a deep learner to detect face identity manipulation. However, the use of the Xception, a single CNN based model, for deepfakes detection limits its forging artifact understanding capability as discussed earlier. Moreover, the inclusion of the face recognition technique, a VGG-19 based model, makes the model not only heavy resource-oriented but also the generalization capability of the face detector models cannot be assured when it tries to recognize an unknown face image. Recently, Mohiuddin et al. (2022) proposed a feature fusion scheme to minimize the generalization error and dependency on a single CNN based model. In this work, the authors used fine-tuned MesoInception-4 model to extract features from three different color spaces, namely RGB, YCbCr, and HSV and then concatenated them to form a single feature vector. This method was able to minimize the generalization error but not at par with the state-of-the-art method like Nirkin et al. (2021). In another work, Bonettini et al. (2021) studied the effect of an ensemble model when using different forms of a CNN model. In their work, they used EfficientNet (Tan & Le, 2019) as CNN model and average voting as ensemble model. The models were evaluated using intra-dataset framework and thus generalization capability (i.e., performance at inter-dataset setup) of this work cannot be ensured.

In contrast to these methods, researchers (e.g., Chen, Yao et al., 2021, Coccomini et al., 2022, Dang et al., 2020, Ganguly et al., 2022, Heo et al., 2021, Wang & Deng, 2021, and Wodajo & Atnafu, 2021) used either attention based mechanisms (Chen, Yao et al., 2021; Dang et al., 2020; Ganguly et al., 2022; Wang & Deng, 2021) which try to put more focus on certain parts of input images or associated vision transformer (Coccomini et al., 2022; Heo et al., 2021; Wodajo & Atnafu, 2021) that attempted to extract relation among different patch generated from a cropped face image. Dang et al. (2020) used an attention mechanism to concentrate the manipulation region and expand feature representation for face forgery detection. Wang and Deng (2021) designed a method that can detect face forgeries

by an attention-based data augmentation method which guides the detector to refine and enlarge its focus. To exploit local features of the face region, an attention mechanism was applied by Chen, Yao et al. (2021) where information represented by RGB and frequency domain were fused to extend the model's capability of artifact pattern similarity learning. Recently Ganguly et al. (2022) proposed a visual attention based model to carry out detection of deepfakes. This method considered single frame based deepfakes detection for video data. In general, attention-based methods performed better than naive CNN based methods described earlier. However, these methods searched for artifacts all over the image but did not learn correlation among artifacts of different parts of the face regions. Specifically, attention-based methods can activate artifacts present in a part of a face image but fail to extract relation among those artifacts appeared apart (Heo et al., 2021; Wodajo & Atnafu, 2021).

As a result, Coccomini et al. (2022), Heo et al. (2021) and Wodajo and Atnafu (2021) proposed deepfakes detection methods that include ViT to learn correlation among distant artifacts. Wodajo and Atnafu (2021) designed a Convolutional ViT (CViT) which has a similar working principle like the model, known as Convolutions to Vision Transformers (CVT), proposed by Wu et al. (2021). In this CViT model, convolutional layers were designed following VGG model's convolutional architecture to extract feature maps. Next, the extracted feature maps from input face images which were fed to ViT model to learn the correlation among the feature maps. Whereas, Heo et al. (2021) introduced an improved version of the CViT architecture. In this work, the authors concatenated the features extracted from entire face image using a pre-trained EfficientNet model (rearranged to the number of segments equal to number of patches) with the patch-wise features generated using learnable embedding information before passing to the ViT architecture. The feature fusion was performed using global average pooling scheme. Chronologically, the latest ViT-inspired method is Convolutional Cross ViT (CCViT) model proposed by Coccomini et al. (2022). CCViT used two distinct branches, namely S-branch and L-branch similar to CViT. The former considered small-sized patches (7×7), and the later used large-sized patches (54×54 for EfficientNet, and 64×64 for convolutional architecture used in CViT) to obtain limited and wider receptive fields respectively. Coccomini et al. (2022) and Heo et al. (2021) used EfficientNet (Tan & Le, 2019) as backbone CNN model. This choice was inspired from the fact that this CNN model provided the best performance among the state-of-the-art CNN models on the DFDC dataset. Hence, performances of these models on inter-dataset experimental setup cannot be inferred. Moreover, these ViT-inspired models were trained on a large volume of facial images extracted from video frames which is also a requirement of ViT model to perform satisfactorily. However, training of such an architecture on voluminous data needs heavy computational resources which might not be available to the general user. In other words, results obtained by these models cannot be generalized while trained with lesser training samples or samples generated with less variety of deepfaking methods as compared to target the dataset.

3. Motivation and novelties

Detection of deepfakes multimedia content is a challenging but contemporary research problem. As a result, many researchers have contributed a number of solutions to solve different issues related to the detection of deepfakes in the recent past. Additionally, a number of large-sized datasets have been developed and made available to the research community publicly by several researchers. But the media contents and the number of forging methods used in those datasets are limited. Performances of the state-of-the-art methods (Ganguly et al., 2022; Tan & Le, 2019; Wodajo & Atnafu, 2021; Yu et al., 2022) employing deep learning paradigms are satisfactory when trained and tested on the same dataset but they failed to perform satisfactorily in many cases when trained on one dataset and tested on another (Chen et al., 2022; Li et al., 2020; Mohiuddin et al., 2022). This means these state-of-the-art models face generalization problems. As a result, an experimental trend is found in the literature which considers training samples from different datasets (Coccomini et al., 2022). However, this approach might not eradicate the generalization problem as forgers also try to come up with new faking methods now and then. Thus, it is important to design a system that can work satisfactorily in both intra-dataset (i.e., trained and tested on the same dataset) and inter-datasets (i.e., trained and tested on different datasets) scenarios. This motivates us to design the proposed ViXNet model that can handle both cases. Two parallel feature extraction branches are employed in ViXNet and outputs from these two branches are concatenated at the end to obtain the final features. One branch extracts features from the entire image using a fine-tuned Xception model which we call global features. The other branch utilizes the ViT network to learn correlations among the information obtained from local regions which are here masked patches. The masked patches are generated using a simple self-attention technique which helps in pinpointing the important features. In contrast to other attention based methods (Chen, Yao et al., 2021; Dang et al., 2020; Ganguly et al., 2022; Wang & Deng, 2021), the present attention mechanism is simple and lightweight. Thus, this branch helps in extracting the highly correlated pruned local information present in masked patches.

The use of the ViT network for deepfakes detection could be found in some recent research articles (Coccomini et al., 2022; Heo et al., 2021; Wodajo & Atnafu, 2021) where instead of image patches, feature maps generated by some CNN models (e.g., Heo et al., 2021; Wodajo & Atnafu, 2021) were fed to the ViT model (Dosovitskiy et al., 2021) while Coccomini et al. (2022) fused feature maps generated by a CNN model with the image patches generated using a self-attention network before feeding to ViT model. In contrast to the methods proposed by Heo et al. (2021) and Wodajo and Atnafu (2021), our work uses masked patches as input to the ViT network and thus utilizes advantages of the ViT: dynamic attention, global context fusion, and better generalization (Wu et al., 2021). We have generated masked patches using a self-attention model like Coccomini et al. (2022) but we do not fuse these patches with the learned features which lessen the overhead imposed by the learned features in the ViT. It is worth mentioning that we also use two branches like the work (Coccomini et al., 2022) but in a different way. They used the branches like we do but prior to feeding the information to the ViT network while in our case, the features extracted from the fine-tuned Xception model are stacked with the outputs from the ViT model. By doing so, we are able to use the benefits of the CNN model like local receptive fields, shared weights, and spatial subsampling in a better way. Even keeping these two feature extractor networks in parallel can overcome the failure of one path over the existing ViT-inspired models. Apart from these, we have made extensive experiments to test the performance of the ViXNet in the inter-dataset scenario over existing methods like Coccomini et al. (2022), Li et al. (2020) and Mohiuddin et al. (2022) where a very limited number of such investigations were made.

4. Proposed work

In this work, we propose a deep learning based classification model, called ViXNet, which discriminates forged videos consisting of manipulated faces, generated by different face swapping methods, from the real ones. We approach the solution to this problem by exploiting the inconsistent presence of almost imperceptible leftover artifacts in specific facial regions of manipulated faces. We design a model which can detect the presence of these artifacts by detecting the inconsistencies among the generated image region specific masked features (local feature) coupled with global features of the target face. We achieve this by fusing features from two deep learning models — one model splits the image into patches first and then applies patch-wise soft attention followed by the ViT (Dosovitskiy et al., 2021) to obtain

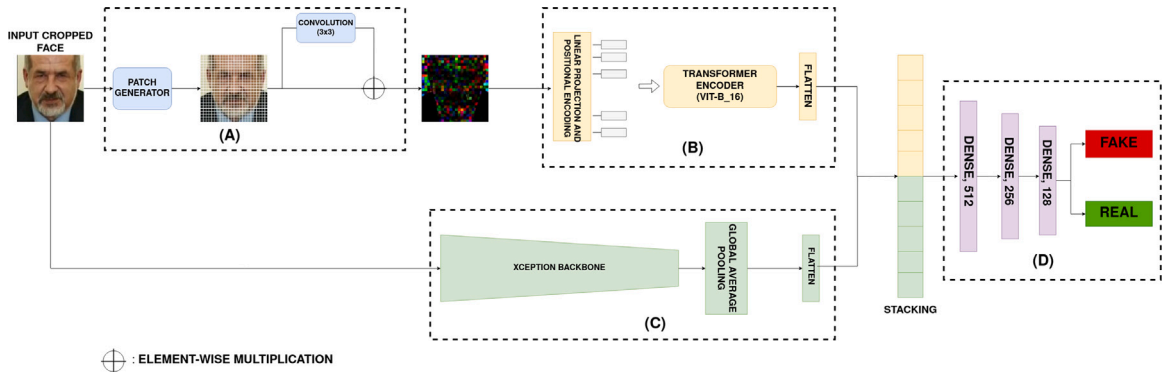


Fig. 2. Proposed ViXNet model which extracts local as well as global image features for deepfakes video detection. It primarily consists of two main components — one of those extracts masked local features first and then tries to capture global inconsistencies among these local features, and another one generates the global image features. (A) Patch-wise self-attention module to generate local features by masking each generated image patch. (B) Multi-headed self-attention module to capture global inconsistencies among the generated local features. (C) CNN backbone to generate global image features. (D) Feature stacking and classification module which is used to concatenate intra-inter image patch features and spatial global features generated from (B) and (C) respectively, and finally identify if the image is either real or fake.

inconsistencies of the masked patch, and another model comprises a deep CNN model, Xception (Chollet, 2017), to generate global spatial features. The working principle of the proposed model is summarized pictorially in Fig. 2. Further descriptions of different components of ViXNet are given in the following subsections.

4.1. Patch-wise self-attention module

This is the first module of ViXNet that takes an input as target image of dimension $(N \times N)$, and splits it down to M patches of size $(P \times P)$ where $\mathbf{M} = (\mathbf{N}/\mathbf{P})^2$. Next, each patch is fed to a patch-wise self-attention module, which primarily consists of a masking mechanism, where each patch $\mathbf{I}$ is multiplied element-wise with a weight matrix $\mathbf{W}$ of dimension $(P \times P)$ to obtain masked patch $\mathbf{J}$ as depicted in Eq. (1). We generate the weight matrix $\mathbf{W}$ by performing (3×3) convolution on the patch $\mathbf{I}$ as shown in block A of Fig. 2.

$$\mathbf{J}_{x:x+P,y:y+P} = (\mathbf{W}_{i,j} * \mathbf{I}_{i,j}), \quad 0 \le x, y \le \mathbf{N} - \mathbf{P}, \quad x \le i \le x + \mathbf{P}, \quad y \le j \le y + \mathbf{P} \tag{1}$$

In this module, each patch is masked with a unique set of trainable weights depending on the location of that patch in the image. Hence, the mask is learned differently for different regions of the image represented by patches of the image. This helps deal with the artifacts present in some specific local regions of the forged face in the image. After obtaining the masked map of the image, we pass it to a transformer for capturing global inconsistencies among these masked patches.

4.2. Global self-attention module

In this module, we intend to locate the inconsistencies among the masked patches obtained from the previous module. This requires computing the similarity of regions not only in local neighborhoods but also the similarity of those regions that lie apart from each other. Inspired by the use of transformers in capturing long-contextual information, we use a transformer for capturing relation among inconsistent artifacts localized in the masked patches. It is composed of multiple transformer blocks. Each transformer block can be broken into a multi-head self-attention block and a feed forward multilayer perceptron. The multi-head self-attention block tries to generate feature maps depicting inconsistencies among the patches. It is achieved by converting the masked image patches, received from the previous module, into linear patch embeddings $X \in \mathbf{R}^{M \times d}$, where $\mathbf{M}$ is the number of patches in the image and $\mathbf{d}$ is the embedding size. Further positional encoding is applied on the embedding space to keep track of the alignment of the patches. The final embedding matrix is projected using three matrices $W_Q \in \mathbf{R}^{d \times d_Q}$, $W_K \in \mathbf{R}^{d \times d_K}$, and $W_V \in \mathbf{R}^{d \times d_V}$ to extract feature representation as Key (say, $\mathbf{K}$), query (say, $\mathbf{Q}$), and value (say, $\mathbf{V}$) which is depicted in Eq. (2), with $d_K = d_Q$ in case of self-attention.

$$Q = XW_Q, \quad K = XW_K, \quad V = XW_V \tag{2}$$

Obtained $\mathbf{K}$ and $\mathbf{Q}$ are used to generate the attention score map, which is a set of weights to be applied on the $\mathbf{V}$ matrix to obtain a feature map depicting only relevant information of the transformed image feature. The attention score map is determined by matrix multiplication of $\mathbf{Q}$ and $\mathbf{K}$ transpose, followed by a scale-down operation by d_k times. Further softmax operator is applied on the score matrix. The obtained score matrix is multiplied with generated $\mathbf{V}$ matrix to form attention output (say, $\mathbf{A}$) as shown in figure Fig. 3. Mathematically, it can be depicted as:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V} \tag{3}$$

In this work, we use a pre-trained transformer model, ViT-B_16. It consists of a spatial embedding block followed by position encoding. ViT-B_16 consists of 12 transformer blocks which make the architecture lighter than other pre-trained architectures like ViT-L_16 or ViT-L_32. Also, the dimension of the patch used in the transformer blocks in case of ViT-B_16 is 16×16. This smaller patch size compared to ViT-B_32, which uses a patch size of 32×32, helps in developing a model of overall light architecture. The model is trained on the ImageNet 2012, which confirms the model learns diverse features compared to model trained on CIFAR-10 or CIFAR-100 dataset. This module returns a feature map which is further fused with a feature map generated by a backbone CNN as described in Section 4.3.

4.3. Global image feature extraction

Xception or Extreme Inception is a deep CNN model that involves linear stacking of depthwise separable convolutions with residual connections. The depthwise separable convolutions deal with both spatial and depth dimensions. The depth dimension initially consists of color channels of an image but increases rapidly with convolutions on the image. In Xception, inception modules of the Inception model are replaced with depthwise separable convolutions. The hypothesis is that the mapping of cross-channel correlations and spatial correlations in the feature map of CNN could be decoupled. Xception has a feature extraction base consisting of 36 convolutional layers, which are structured into 14 modules, as shown in Fig. 4. The Xception architecture has the same parameter count as the Inception architecture, but it uses the parameters more efficiently. It extracts features from different spatial scales and depth scales to generate a final feature map depicting global

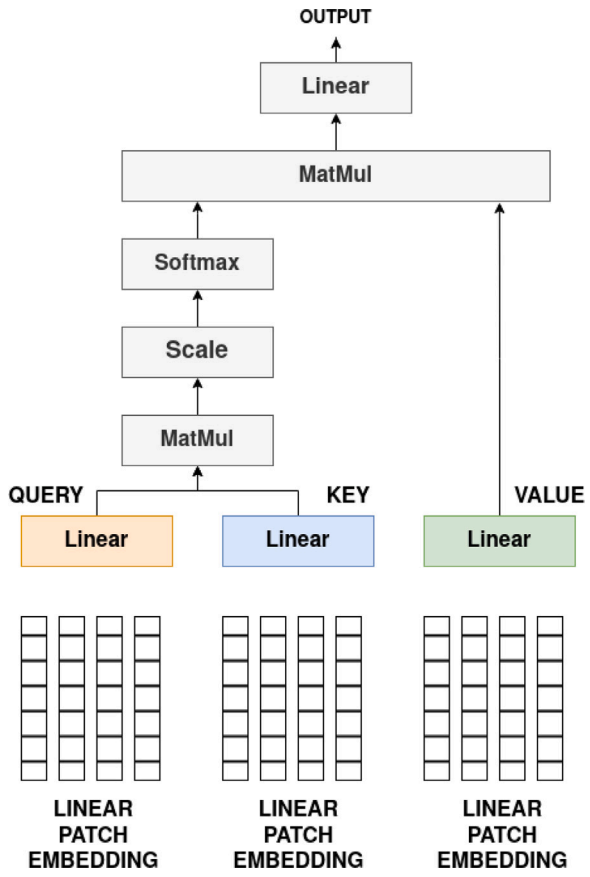


Fig. 3. Architecture of a one-headed self-attention units in transformer blocks, which are helpful in finding global correlation among patch elements. When these units are stacked in parallel to each other, they form multi-headed self-attention units.

features at the end of the network. All these make the Xception model a strong feature extractor and as a result of this, it has been widely used in many computer vision applications. It has also been observed from the literature that it produces state-of-the-art prediction accuracy while this model is trained and tested on the same dataset i.e., in our case the forging artifacts in test samples are known as during training the model, the same artifacts are used. All these motivate us to use this model as a global feature extractor in our work.

4.4. Classification model

From the models described in Sections 4.1, 4.2, and 4.3, we obtain two different types of features, one represents global inconsistencies among the masked patches that help generate intra-inter patch information, and the other one represents spatial global features of the image. In this stage, we first stack two types of features and then design a classifier on the top of the stacked features (see Module (D) in Fig. 2). It is followed by a set of fully connected layers of nodes 512, 256, and 128 respectively which are used to learn a non-linear function over the stacked features. Rectified Linear Unit (ReLU) is used as a non-linear activation function in these layers. For the final classification layer, we use the softmax activation function which returns the probability of an image as fake or real.

5. Experimental results and discussion

It has already been mentioned that in the present work we introduce a deep learning model to determine whether a questioned video contains manipulated face prepared using deepfaking technique or not. In this section, we first introduce details of the forensics datasets used here for experimentation and then describe how train and test datasets are prepared from them. We further discuss the training protocols used. Also, in order to evaluate the model, we test its performance following intra- and inter-dataset experimental setups on three publicly available deepfakes datasets. Such setups help in validating robustness and generalizability of the proposed ViXNet model. Finally, we compare the performance of the proposed model with some state-of-the-art methods. The proposed model is developed in Python programming language using Tensorflow framework. All the experiments have been performed on the Google Colab platform which provides the Nvidia Tesla K80 GPU consisting of 4992 CUDA cores.

5.1. Dataset description

To conduct experiments, we choose three popular, standard and publicly available deepfakes datasets namely, FaceForensics++ (FF++) (Rossler et al., 2019), Celeb-DF (V2) (CeDF) (Li et al., 2020), and Deepfakes Image dataset (DFID) (Afchar et al., 2018). Some sample images/cropped face images from video frame for each dataset are shown in Fig. 5.

FF++: It is a forensics video dataset consisting of 1000 pristine videos that were prepared from 977 Youtube videos by Rossler et al. (2019). All videos contain a trackable face, mostly frontal, without occlusion which enables automated synthetic tampering method to generate realistic forgeries. Next, four automated face manipulation methods — Deepfakes, Face2Face, FaceSwap, and NeuralTextures were employed on each of the pristine category videos to prepare fake videos. Hence this video dataset contains 5000 videos (1000 real and 4000 forged). Since most of the simulated videos available on the web are compressed, hence videos of this dataset were made available in two different quality levels of H.264 encoding. Specifically, they are C-23 that denotes high-quality (HQ) videos and C-40 which represents low-quality (LQ) videos.

CeDF: In this dataset, fake videos were generated by an improved deepfakes manipulation method. It consists of 590 pristine videos and 5369 deepfakes manipulation videos having more than 2.3 million frames in total. The average length of all videos is approximately 13 s with standard frame rate of 30 frames-per-second. The real (i.e., pristine category) videos were collected by Li et al. (2020) from publicly available YouTube videos, corresponding to 59 celebrities with a diverse distribution.

DFID: In addition to the two video datasets discussed above, we also consider the deepfakes image dataset termed here as DFID. It consists of 8000 forged and 11,809 real face images. These were generated from 175 videos collected from different platforms by Afchar et al. (2018). All videos were compressed using H.264 codec with different compression levels, which makes them very close to real images. This dataset consists of only cropped face images from video frames that were extracted by Viola–Jones detector (Viola & Jones, 2001). In order to balance the distribution of face images, the number of frames selected for face image extraction per video is set to be proportional to the number of camera angles and illumination changes on the target face.

5.2. Data preparation

Among the datasets described above, FF++ and CeDF datasets contain videos while DFID contains cropped face images. Therefore, we pass the videos of the first two datasets through a pre-processing step. In this step, we first extract the frames from the videos and then employ multi-task cascading neural network (MTCNN) (Zhang et al., 2016) to detect the most dominant face region in the extracted frames. Finally, cropped face images are extracted from these detected face regions with a 10% padding around the detected region.

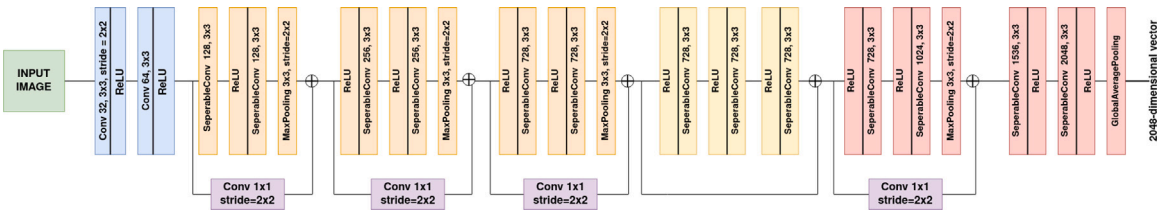


Fig. 4. Architecture of the Xception model, which is used to extract the global feature representation of an input cropped face image.

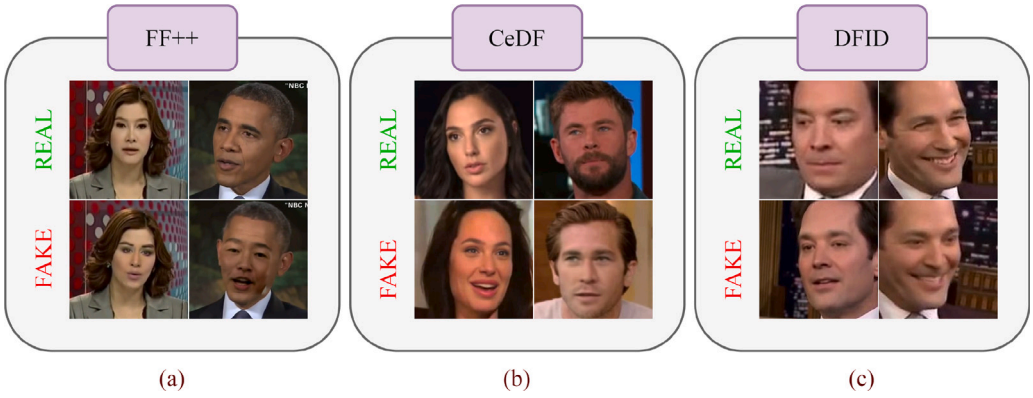


Fig. 5. Some examples of real and forged images (generated by face swapping techniques) taken from the datasets — (a) FF++, (b) CeDF, and (c) DFID.

In FF++ dataset, we use only Deepfakes subsets, owing to our focus on detection of videos which are forged by face swapping techniques. We divide the 1000 videos of each class (Deepfakes and Pristine) into two classes where each class is divided into three subsets — 70% as the training set, 20% as the validation set, and 10% as the test set. We generate 5876 cropped face images from 1400 of the training videos, 400 cropped images from 400 validation videos, and 200 cropped face images from test videos in total. While generating the training image set, equidistant frames in the temporal space are chosen. This gives a variety of face postures in the image set because no two frames would be adjacent to each other. While generating the validation and test image sets, only the first frame of the videos is selected.

In the case of the CeDF dataset, we use 5011 videos for generating the training image set, 1000 videos for generating the validation image set, and 518 videos for generating the test image set. It is noteworthy in mentioning that the test videos are the videos that were presented by the dataset providers. Cropped faces for the train, validation and test sets are generated similar to the FF++ case. Thus, we obtain 9155, 1000 and 518 train, validation and test samples respectively. The DFID consists of 19 809 images, in which 8000 are forged images, and 11 809 real images. We use 11 114 images as training samples, 1238 as validation samples, and 7104 as test samples. The test samples are samples that were used in Afchar et al. (2018). To have better cropped face images, all these samples are passed through the present face cropping technique. In Table 1, we summarize the data distribution used for model evaluation. It is to be noted here that for videos we have used single frame based detection technique like Ganguly et al. (2022) and Mohiuddin et al. (2022) i.e., we have used only one frame from a test video to decide whether the video is fake or real.

Table 1
Distribution of generated frames used in training, validation and test image sets of each dataset.

Dataset	#Real frames in			#Fake frames in		
	Training	Validation	Test	Training	Validation	Test
CeDF	1130	100	178	8022	900	340
FF++	2930	200	100	2946	200	100
DFID	6549	700	4259	4565	538	2845

5.3. Experimental setups

It has already been discussed that a good forensic detection model can distinguish forged cases from real ones irrespective of datasets and forging algorithms used. In this work, we try to achieve this ability by the proposed ViXNet model. To test how generalizable or robust the ViXNet model is, we have experimented it with two different approaches — intra-dataset approach and inter-dataset approach. In the first case, ViXNet is trained, validated, tested on a single dataset, while in the second case it is trained, validated and tested on different datasets. To be more specific, in case of intra-dataset experimentation, we train ViXNet on a set say A_{train}, and select the best model based on the set $A_{validation}$, and further test the best model on a set A_{test}. With this procedure, we train a model which works best on a given dataset $A_{validation}$. In case of inter-dataset experiments, ViXNet is trained on a set say A_{train}, and select the best model based on the set $B_{validation}$, and further test the best model on a set C_{test}. In this procedure, the model is evaluated on a different dataset than while training, which results in a hybrid model capable of detecting artifacts from both datasets. We discuss the results generated by these experiments in the later sections. Since the forensics datasets — FF++, CeDF, and DFID are generated by different face swapping techniques thus, performances of ViXNet in inter-dataset experimental approach will describe how effective this model is irrespective of the dataset used during model training.

5.4. Model parameters

In this work, we train ViXNet as a whole, despite it consisting of different trainable components. The backbone CNN, i.e., the Xception model, used for generating global features is initialized with weights trained on the ImageNet dataset. We use ViT-B_16 as the global self-attention module, which is pretrained on the ImageNet dataset. Other trainable components like the patch-wise self-attention module and the classification module are initialized with random weights. During training and testing of the model, we use images resized to a fixed dimension of 384 × 384. To train the model, we set the number of

Table 2
Performance of ViXNet and combinations of its different components on an intra-dataset experimental setup, where the model is trained, evaluated, and tested on DFID.

Model	Test accuracy (%)	AUC score (%)
PWSA	71.02	71.99
ViT	59.95	67.28
Xception	95.20	98.80
PWSA+ViT	59.95	69.69
ViT+Xception	95.95	98.89
PWSA+Xception	83.47	92.85
ViXNet	95.42	98.93

epochs to 50, and a learning rate of 0.0001. A small learning rate enables the model to take small steps while reducing the loss, thereby ensuring the model convergence to an optimal solution. The binary cross entropy has been used as a loss function as the classification output is binary — real or fake. We use Adam as the optimizer while training. During model training and validation, we set the batch size to 8 and steps-per-epoch to 40 and 50 respectively.

5.5. Results using intra-dataset experimental setup

In this section, we discuss the performances of ViXNet itself and combinations of its different components by using an intra-dataset experimental setup. It helps us to understand how the entire model or its components perform when the test artifacts are known during model training. Here, we evaluate seven different models (entire model and six combinations of its components) which are as follows.

1. Model having only patch-wise self-attention module, hereafter termed as PWSA. In this model, the learned patch-wise masks are multiplied with corresponding patches to generate local features.
2. Model consisting only pre-trained ViT model, hereafter called as ViT. Here, the original image patches are fed to the pre-trained ViT model for learning global correlation among the original image patches.
3. Model made with only pre-trained Xception model, hereafter termed as Xception. Here, we directly feed the cropped faces to the exception model to obtain the classified result based on global spatial image features.
4. Model prepared by keeping PWSA and ViT i.e., (PSWA+ViT). Here, we try to generate features which depict correlations among masked local facial features.
5. Model prepared by stacking features from PWSA and Xception i.e., PSWA+Xception. Here, we try to generate features based on spatial structure of locally masked cropped face images.
6. Model prepared by stacking features from ViT and Xception (ViT+Xception). In this case, we try to generate features resembling global correlations among various patches of face regions and feature maps obtained by spatial feature extraction by Xception.
7. The entire model i.e., ViXNet.

For simplicity, we report performances of these seven models in terms of test accuracy and AUC score for DFID (see Table 2). This table shows that the performance of the Xception model is very close to the overall model performance. Also, use of either ViT or PWSA with the Xception model reduces the performance of the Xception model. However, when both are coupled with the Xception model (i.e., ViXNet), the performance gets improved. Overall, ViXNet performs slightly better than the Xception model, as discussed earlier, when train and test set images have almost similar artifacts in fake faces. However, our model has a better ability to learn imperceptible artifacts than that of the Xception model.

Performances of ViXNet on each of the datasets are recorded in Fig. 6 while the corresponding confusion matrices are shown in Fig. 7. In Fig. 8, we present various plots depicting learning curves of the proposed model. We obtain nearly monotonous loss and accuracy curves, which also indicates that the model is properly trained.

5.6. Results using inter-dataset experimental setup

Here, we discuss the performance of ViXNet and its different components (described in Section 5.5) on inter-dataset experimental setup. We perform training, validation and testing on different datasets to evaluate the generalizability of the model. We report the results generated in Table 4. We follow experimentation protocols as discussed in Section 5.3. It is to be noted that for simplicity, we only provide the results of ViXNet and combinations of its different components while trained on DFID in Table 3. It is evident that the proposed model ViXNet outperforms almost each of its combinations of components in terms of both test accuracy and AUC score. ViXNet when tested on FF++ performs well in generalizing the classification results, outperforming all of its combinations of components. However, when tested on CeDF dataset, the performance is relatively less.

We also evaluate the models trained on a training set of one dataset and tested on a test set of other datasets. The results are recorded in Table 4. These results show that the proposed model performs well in all of the cases. Fig. 9 shows the confusion matrices obtained by evaluating the ViXNet following an intra-dataset experimental setup. It can be observed that most of the misclassifications occurred during testing on FF++ (see Figs. 9(a) and 9(f)) and the minimum number of misclassifications happens during testing on CeDF dataset (see Figs. 9(b) and 9(d)).

5.7. Comparison with state-of-the-art methods

In this section, we compare the performance of proposed model ViXNet with some state-of-the-art deepfakes detection methods proposed by Afchar et al. (2018), Chollet (2017), Coccomini et al. (2022), Ganguly et al. (2022), Guo et al. (2021), Heo et al. (2021), Mohiuddin et al. (2022), Qian et al. (2020), and Wodajo and Atnafu (2021). For fair comparison, we have evaluated these methods following our experimental protocols as described earlier. This would help in comparison of different methods based on factors like robustness and generalizability. We report the performances in terms of test accuracy and AUC score of the said deepfakes detection methods on our experimental setup (see Section 5.3), on FaceForensics image dataset, in Tables 5 and 6 respectively. The first column of the table Tables 5 and 6 states the experimental protocol used here while evaluating the performance of the methods. The naming of the protocols goes by the convention $Dataset_{train}$–$Dataset_{validation}$–$Dataset_{test}$. It can be observed that ViXNet outperforms all the state-of-the-art methods used here for comparison in case of intra-dataset experimental setup. In case of inter-dataset experiments, which is basically an evaluation of generalizability of a model, ViXNet outperforms and achieves state-of-the-art results in most of the experiments.

The comparative results shown in Table 6 can be interpreted as follows: (1) It can be observed that ViXNet which uses Xception as the global spatial feature extractor outperforms base Xception network in all of the inter-dataset experimental protocols which confirms the introduction of Patch Wise Self-Attention module and Global Self-Attention module has positive contributions to the base CNN model in generalization tasks. (2) It can be observed that ViXNet outperforms F3-Net which primarily takes frequency information into consideration for face forgery classification. Outperforming the network ensures the capability of the proposed ViXNet model to handle noisy input image information caused due to image compression algorithms. (3) It can also be observed that ViXNet outperforms the AMTEN network in every experimental protocol. AMTEN used a similar approach of masking the original image to suppress the original image content. However, the

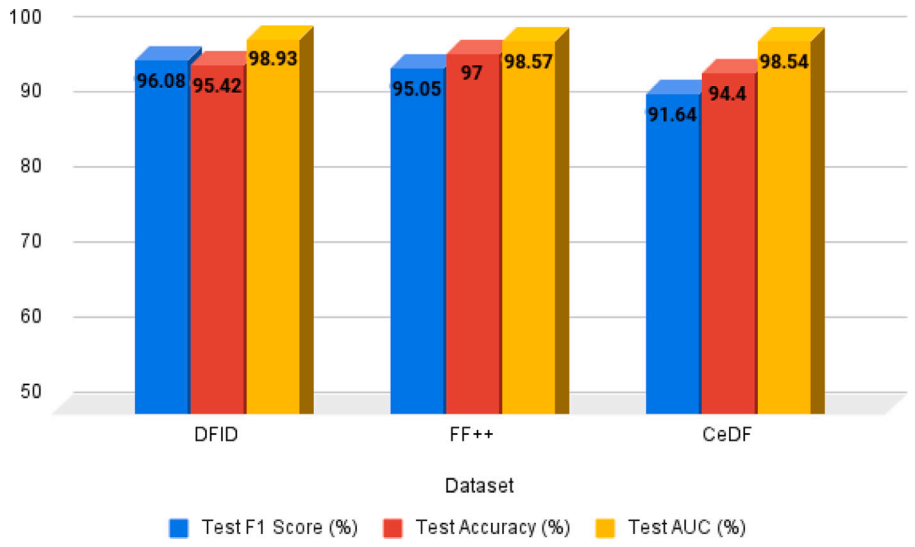


Fig. 6. Performance of ViXNet on the three datasets used here when intra-dataset experimental setup is used.

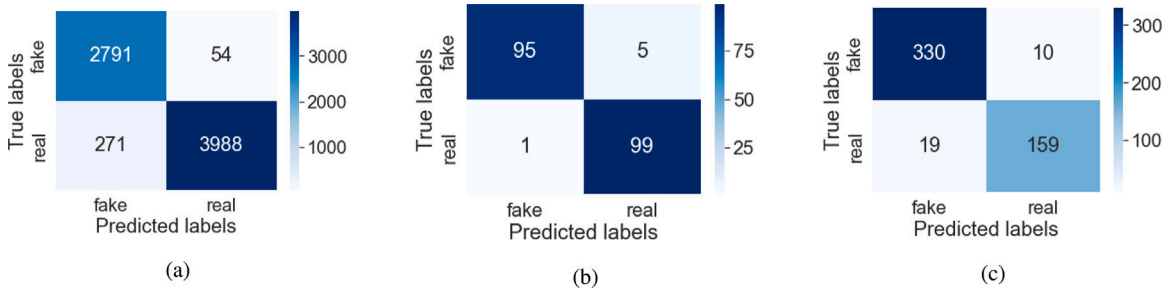


Fig. 7. Confusion matrices of the intra-dataset experimental setup — ViXNet is trained, validated and tested on (a) DFID, (b) FF++, and (c) CeDF datasets.

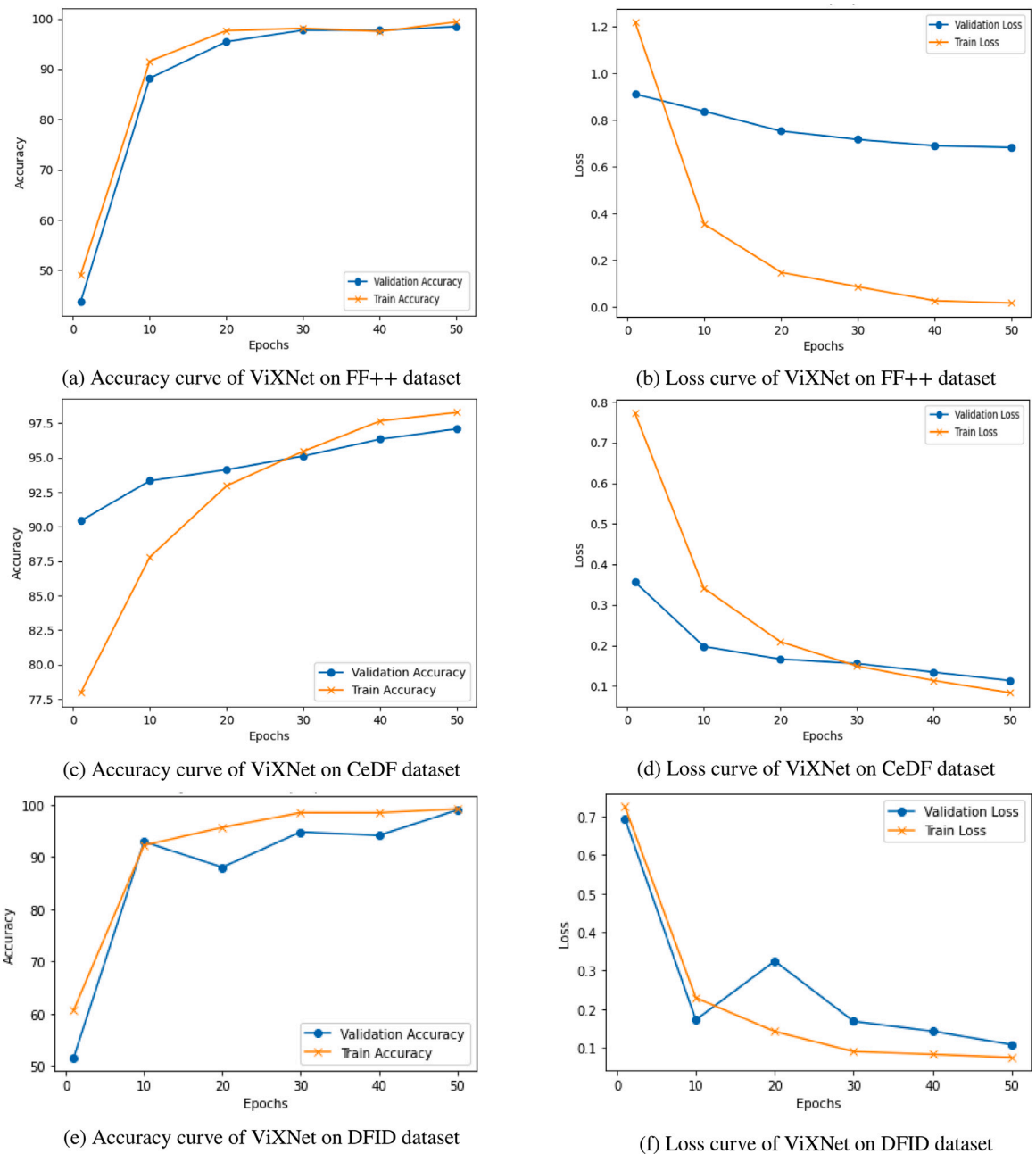
Table 3
Performance of ViXNet and combinations of its different components on inter-dataset experimental setup. The results shown here are obtained after training on DFID only.

Validation dataset	Test dataset	Model	Test accuracy (in %)	AUC score (in %)
CeDF	FF++	PWSA	41.70	40.90
		ViT	50.75	53.19
		Xception	66.83	79.18
		PWSA+ViT	50.75	53.18
		ViT+Xception	61.30	69.48
		PWSA+Xception	68.34	75.18
		ViXNet	75.37	83.60
FF++	CeDF	PWSA	50.25	46.82
		ViT	34.36	34.53
		Xception	56.17	58.29
		PWSA+ViT	34.36	34.52
		ViT+Xception	59.84	64.02
		PWSA+Xception	68.53	70.64
		ViXNet	61.19	66.36

Table 4
Performance of ViXNet on the three datasets used, when inter-dataset experimental protocols are followed.

Dataset			Performance (in %) on test set				
Train	Validation	Test	Accuracy	AUC score	Recall	Precision	F1-score
DFID	CeDF	FF++	75.50	83.60	97.00	68.00	79.95
DFID	FF++	CeDF	61.19	66.36	70.22	45.78	55.43
FF++	CeDF	DFID	68.90	75.13	55.69	76.46	64.44
FF++	DFID	CeDF	69.30	74.78	45.45	39.21	41.88
CeDF	FF++	DFID	65.08	71.76	85.53	64.13	73.33
CeDF	DFID	FF++	68.00	75.22	69.00	71.00	69.99

masking process was done globally, with no region-specific masking operations. Outperforming the network implies that local region-specific masking process by PWSA module helps in improving the classification performance.(4) In contrast to our method, the Convolutional-VIT uses a deep CNN and a vision transformer sequentially, taking the output of the CNN as input to the VIT. Since our method outperforms this method, we can see the advantages of taking an independent approach, where the CNN and VIT are decoupled and jointly work to extract features at different scales. (5) The method utilized by Coccomini et al. (2022) also extracts features at different scales by extracting patches of two different sizes. Since this method also performs quite robustly, we see that extracting information at different scales for deepfakes detection is effective.



(a) Accuracy curve of ViXNet on FF++ dataset

(b) Loss curve of ViXNet on FF++ dataset

(c) Accuracy curve of ViXNet on CeDF dataset

(d) Loss curve of ViXNet on CeDF dataset

(e) Accuracy curve of ViXNet on DFID dataset

(f) Loss curve of ViXNet on DFID dataset

Fig. 8. Plots of various metrics obtained while training ViXNet on various datasets following an intra-dataset experimental protocol.

Table 5
Comparison of test accuracy (in %) of the proposed model with various state-of-the-art models evaluated on our experimental setups. The naming of the experimental protocols goes by the convention Train Dataset_Validation Dataset_Test Dataset. In this table, M1–M9 represent the methods proposed by Afchar et al. (2018), Afchar et al. (2018), Chollet (2017), Coccomini et al. (2022), Ganguly et al. (2022), Guo et al. (2021), Mohiuddin et al. (2022), Qian et al. (2020) and Wodajo and Atnafu (2021) respectively.

Experimental protocol	M1	M2	M3	M4	M5	M6	M7	M8	M9	ViXNet
FF++_FF++_FF++	67.00	65.00	95.50	95.50	78.00	86.00	86.00	84.00	96.00	97.00
CeDF_CeDF_CeDF	65.64	65.83	89.38	87.06	68.33	65.64	84.56	81.32	90.35	94.40
DFID_DFID_DFID	80.28	78.91	95.21	95.05	84.50	84.71	88.37	82.57	94.76	95.42
FF++_CeDF_DFID	57.28	53.74	71.02	72.71	41.82	62.94	62.53	66.47	63.27	68.90
FF++_DFID_CeDF	51.54	66.41	55.60	63.89	41.31	44.21	63.71	68.04	62.36	69.30
CeDF_FF++_DFID	40.51	45.34	60.36	56.39	57.34	40.05	60.84	67.25	62.80	65.08
CeDF_DFID_FF++	70.50	50.00	65.00	54.00	64.00	50.00	60.00	65.00	68.00	68.00
DFID_CeDF_FF++	63.00	66.00	67.00	76.00	63.00	65.00	66.00	60.00	57.00	75.50
DFID_FF++_CeDF	59.85	64.48	56.17	62.35	55.79	61.78	63.13	59.40	47.30	61.19

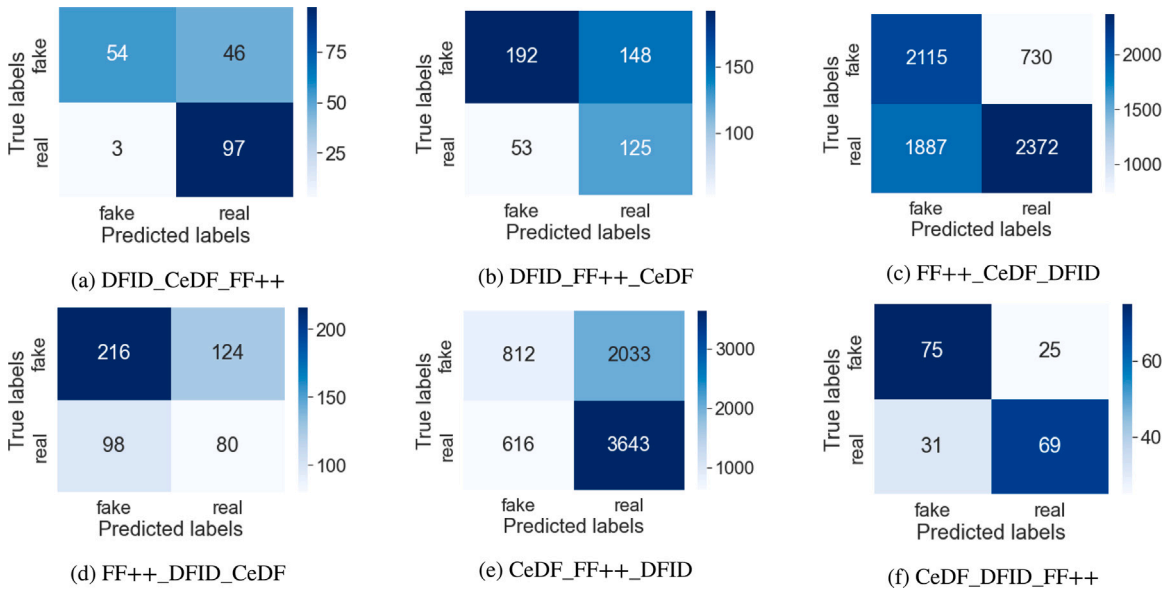


Fig. 9. Confusion matrices of different inter-dataset experimental setups. Here, X_Y_Z indicates that ViXNet is trained on dataset X, validated on dataset Y and evaluated on dataset Z.

Table 6
Comparison of test AUC scores (in %) of the proposed model with various state-of-the-art models evaluated on our experimental setups. The naming of the experimental protocols goes by the convention Train Datase_Validation Dataset_Test Dataset. In this table, M1–M9 represent the methods proposed by Afchar et al. (2018), Afchar et al. (2018), Chollet (2017), Coccomini et al. (2022), Ganguly et al. (2022), Guo et al. (2021), Mohiuddin et al. (2022), Qian et al. (2020) and Wodajo and Atnafu (2021) respectively.

Experimental protocol	M1	M2	M3	M4	M5	M6	M7	M8	M9	ViXNet
FF++_FF++_FF++	66.81	66.93	98.79	96.52	87.62	96.41	85.92	83.04	99.06	98.57
CeDF_CeDF_CeDF	65.33	64.80	95.59	81.48	78.04	61.07	78.46	80.50	96.63	98.54
DFID_DFID_DFID	70.33	62.53	98.80	95.70	92.11	92.21	89.12	79.51	98.47	98.93
FF++_CeDF_DFID	56.68	50.36	78.17	68.04	40.60	74.24	59.91	63.56	74.71	75.13
FF++_DFID_CeDF	51.33	65.58	58.06	53.89	40.73	50.96	56.43	66.12	63.28	74.78
CeDF_FF++_DFID	45.86	44.37	67.21	62.17	56.80	47.47	52.82	62.36	62.64	71.76
CeDF_DFID_FF++	69.68	51.51	75.19	54.04	71.09	72.37	59.93	63.80	78.84	75.22
DFID_CeDF_FF++	55.60	62.28	79.18	75.96	63.77	73.22	65.70	57.19	82.37	83.60
DFID_FF++_CeDF	54.17	61.65	58.29	59.14	59.17	50.00	54.24	59.01	60.60	66.36

Table 7
Performance comparison of different models tested on the FaceForensics image benchmark platform. Here, DF, F2F, FS and NT represent DeepFakes, Face2Face, FaceSwap and NeuralTexture respectively.

Method	DF	F2F	FS	NT	Pristine	Total
Fridrich and Kodovsky (2012)	73.60	73.70	68.90	63.30	34.00	51.80
Cozzolino et al. (2017)	85.40	67.80	73.70	78.00	34.40	55.20
Rahmouni et al. (2017)	85.40	64.20	56.30	60.00	50.00	58.10
Bayar and Stamm (2016)	84.50	73.70	82.50	70.60	46.20	61.60
Rossler et al. (2019)	74.50	75.90	70.90	73.30	51.00	62.40
Qian et al. (2020)[a]	76.40	75.20	86.40	86.00	52.80	56.80
Guo et al. (2021)[a]	44.50	37.20	34.00	23.30	77.00	55.50
Chollet (2017)[a]	89.10	88.30	69.90	88.00	43.00	64.00
ViXNet[a]	89.10	78.10	66.00	84.00	60.00	69.90

[a]Methods are evaluated on our experimental setup.

5.8. Results on FaceForensics benchmark image dataset

To test the robustness of ViXNet model, we evaluate its performance on FaceForensics benchmark image dataset which has 1000 unlabeled target test images. Samples of this test set belong to any one of the five classes as found in FF++ dataset - DeepFakes, FaceSwap, Face2Face, NeuralTexture and Pristine. As forged images of this test set are generated by both face swapping and face reenactment methods, a new training image dataset is prepared consisting samples from each of the said classes. A total of 15 000 training images are extracted from the FF++ video dataset which are equally distributed over the five classes. Next, 3000 samples are selected from the train set as validation image samples. We train the model on the generated training image set, and selected the best validated model for benchmark evaluation. In the benchmark images, the results are computed on a private server by uploading predictions in a formatted JSON file. Besides ViXNet, we train some more methods on our generated dataset for evaluating them on the benchmark.

We report the benchmark results in Table 7. The proposed model ViXNet obtains an overall accuracy of 69.90% on the benchmark images. Table 7 shows that our proposed model outperforms most of the previous methods in terms of overall accuracy. The accuracy of the different categories cannot be used to evaluate the model performance, since there is a unequal number of real and fake images in the test

Table 8
Performance comparison of different models tested on the DFDC dataset.

Method	#Training samples	Performance (in %)	
		AUC score	F1-Score
Heo et al. (2021)	>1 000 000	97.80	91.90
Wodajo and Atnafu (2021)	162,174	84.30	77.00
Guo et al. (2021)[a]	86,166	74.19	72.83
Coccomini et al. (2022)	220,444	92.50	84.50
Mohiuddin et al. (2022)[a]	86,166	71.77	68.00
Ganguly et al. (2022)[a]	86,166	83.59	74.21
ViXNet (Single frame based decision)	86,166	86.32	79.06
ViXNet (Multiple (12) frames based decision)	86,166	90.26	83.18

[a]Methods are evaluated on our experimental setup.

image benchmark dataset. However, it is observed our model predicts real images correctly better than most of the existing methods.

5.9. Results on DFDC dataset

First introduced in a Kaggle contest (Dolhansky et al., 2019) and later opened by Facebook AI (Dolhansky et al., 2020), the DFDC dataset is one the most comprehensive third-generation forensics datasets to date. The training set consists of 119,154 ten second clips of 486 subjects. Morphed videos are generated using various models such as DFAE, MM/NN faceswap, NTH, FSGAN, StyleGAN. The validation set consists of 4000 ten second clips of 218 subjects. Finally, there is the private test set consisting of 10 000 clips and the public test set consisting of 5000 clips. We have constructed our train set by picking a random frame from 86,166 videos of the training set, for a total of 86,166 images. Of these 13,916 belong to the real class, while the remaining 72,550 images belong to the fake class. In order to maintain a good balance between these two classes, we have over-sampled the real images during training time.

We have constructed our test set by picking the first frame from the public test set. We have trained our model for 400 epochs, using the Adam optimizer with a learning rate of 0.0001, and as detailed in Table 8 we obtained an AUC score of 86.32% and an F1score of 79.06%. We also evaluate the performance of the model obtained by replacing the XceptionNet backbone by the EfficientNet B0 and obtained AUC score, and F1score as 86.30% and 78.29% respectively. This result infers that the proposed model performs well compared to use of EfficientNet B0 that was used in Coccomini et al. (2022) and Heo et al. (2021) when tested on present setup. We have also evaluated our model following multiple frames based decision approach where at most 12 random frames have been taken from a test video. The final prediction score for a video is obtained by averaging the scores for the associated frames. In this setup, we have obtained better results over when followed the single frame decision approach (see Table 8). We note that due to computational limitations, we could not effectively train on such a large dataset. We have used only one frame per clip for a total 86,166 frames as opposed to the 220,444 by Coccomini et al. (2022), we have also trained with 40 batches per epoch in contrast to 2500 used by Heo et al. (2021). The performances of some state-of-the-art methods along with ViXNet are recorded in Table 8. In order to highlight the generalizability of our model, we have also provided intra-dataset results, where we have trained on DFID, FF++, and CeDF datasets and tested on the DFDC dataset (see Fig. 10).

We also report the performances of the state-of-the-art and ViXNet, in terms of AUC score and F1-score, of the discussed methods on the DFDC dataset in Table 8. It can be observed that the proposed method performs better than most of the state-of-the-art methods, evaluated on our experimental setup. It should be noted that Guo et al. (2021) followed a similar approach of globally masking the original face image. The proposed method outperforms the former, which shows the effectiveness of the local region-specific masking approach. Ganguly et al. (2022) also used an attention based approach, with the Xception network as the backbone. The proposed method also uses the Xception backbone, so it can be confirmed that the region based masking followed by global self-attention outperforms the soft attention used in the former. However, it should be noted that Heo et al. (2021) used feature extraction by deep CNN model coupled with vision transformer and distillation outperforms the proposed method in terms of AUC Score, and F1-Score. It should also be noted that in the discussed methods, every frame of videos in the DFDC video set is used in generating the training set, whereas in our method the dataset is prepared with only one frame per video. Hence, there is a large difference in the number of training image samples used in both the cases, the result of which is reflected upon evaluating the methods on the testing dataset.

5.10. Error analysis

Here, we discuss and analyze the performance of the proposed model ViXNet at the image level. For analysis, we have evaluated the model on the generated test image dataset as discussed in Section 5.2, following intra-dataset experimental setup (see Section 5.3). After evaluating ViXNet on the said test set, we try to understand the cases when our model classifies the images properly, and when it fails. In Fig. 11, we present images which are classified properly (see Figs. 11(a) and 11(b)) as well as wrongly (see Figs. 11(c) and 11(d)). It can be observed from Fig. 11 that the false positives (see Fig. 11(c)) or the fake images classified as real have a common artifact which seems like a change in lighting condition on the face region. It seems that the proposed model is not able to encode this type of artifact in its present version. It is also observed that the false negatives (see Fig. 11(d)) or the real images classified as fake have some facial expressions or facial occlusions, which causes the model to perceive certain expressions as forged artifacts. This supposedly arises due to the lack of certain facial expressions in the training image set. It can also be observed that the proposed model can correctly classify forged images (i.e., true positives) with imperceptible face region-specific artifacts, thereby confirming its superiority over the existing methods.

6. Conclusion

In this paper, a deepfakes image forgery detection network, called ViXNet, has been proposed. The ViXNet can identify a forged face by exploiting the presence of leftover imperceptible artifacts. It uses a patch-wise self-attention module to obtain face-region specific masks and further tries to compute the inconsistencies among these masked regions, coupled with global spatial features generated by a deep CNN network for final classification. Extensive evaluation of the model is performed on two different scenarios — intra-dataset and inter-dataset to quantitatively validate the robustness and generalizability of the model. The proposed method outperforms many state-of-the-art methods in terms of classification accuracy and AUC score. There have been previous methods which involve using local region-specific features in the ViT model. However, they involve extracting high level features of the face image and further passing them into the ViT. In the current work, image

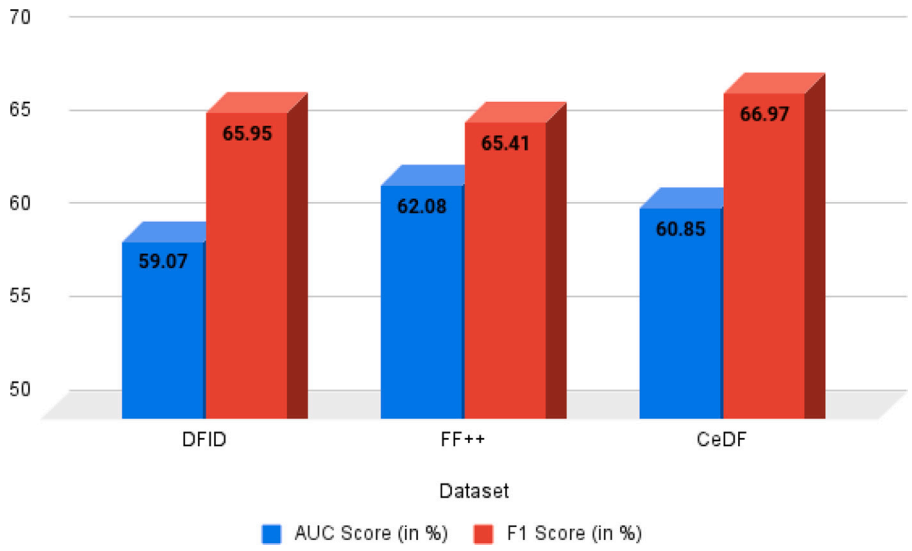


Fig. 10. Performance of ViXNet on DFDC when an intra-dataset experimental setup is used.

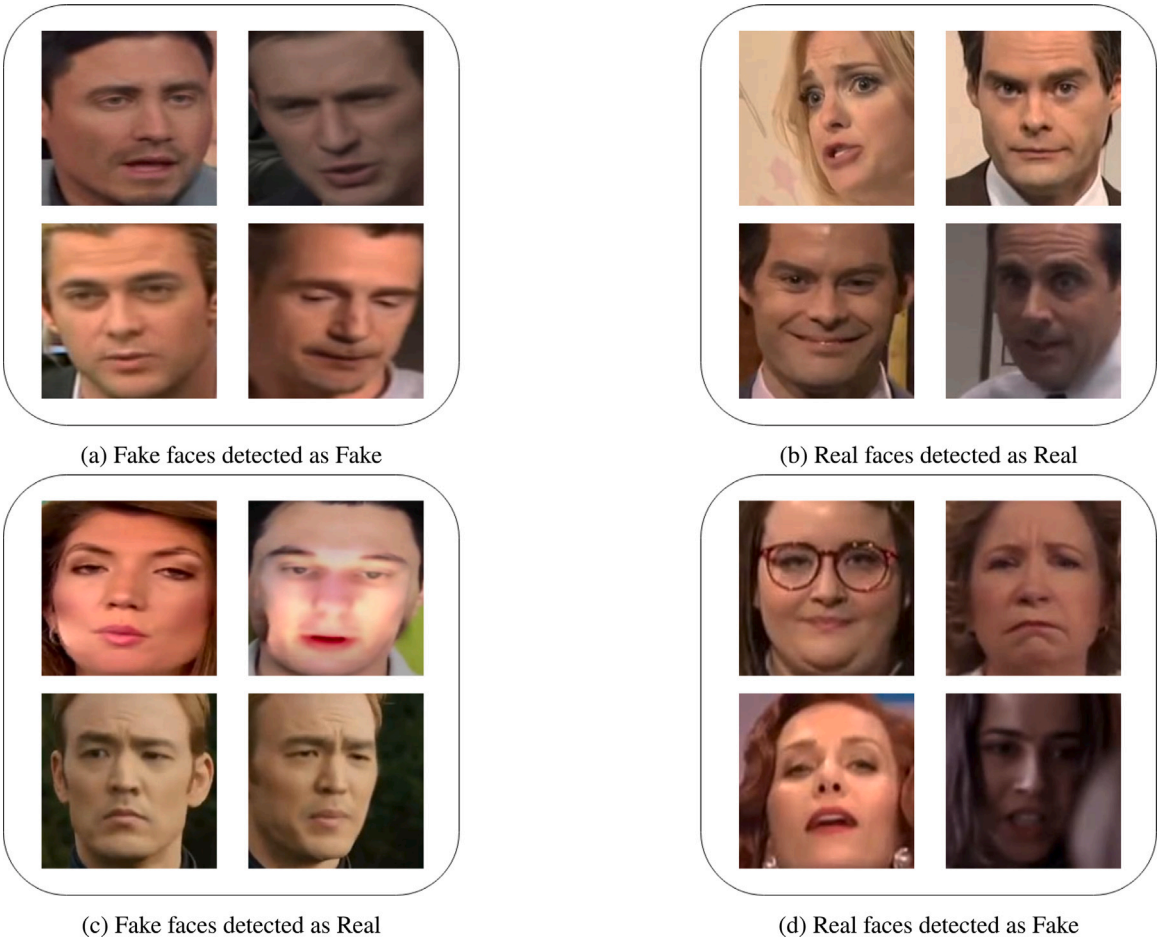


Fig. 11. Some correctly and wrongly classified face images by the proposed model. Here we show samples that are either classified correctly (i.e., (a) true positive and (b) true negative), or erroneously (i.e., (c) false positive and (d) false negative cases).

regions of a specific size are used to generate a region-specific masked image, the correlation among which are computed by a lightweight visual transformer, ViT-B_16. The masking is also performed on a low-level feature selection based approach using the convolution operation. In our future work, a strategic multi-scale approach can be considered, where image regions of varied sizes are to be used for generating masks. Also, a deep visual transformer like ViT-B_32 or ViT-L_32 can be used for learning better inconsistencies among masked image regions.

CRediT authorship contribution statement

Shreyan Ganguly: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, Visualization. **Aditya Ganguly:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, Visualization. **Sk Mohiuddin:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, Visualization. **Samir Malakar:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing, Supervision. **Ram Sarkar:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing, Visualization, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgments

The authors would like to thank the Centre for Microprocessor Applications for Training, Education and Research (CMATER) research laboratory of the Computer Science and Engineering Department, Jadavpur University, Kolkata, India for providing the infrastructural support.

References

Afchar, D., Nozick, V., Yamagishi, J., & Echizen, I. (2018). Mesonet: a compact facial video forgery detection network. In *2018 IEEE international workshop on information forensics and security (WIFS)* (pp. 1–7). IEEE.

Amerini, I., Galteri, L., Caldelli, R., & Del Bimbo, A. (2019). Deepfake video detection through optical flow based cnn. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*.

Bayar, B., & Stamm, M. C. (2016). A deep learning approach to universal image manipulation detection using a new convolutional layer. In *Proceedings of the 4th ACM workshop on information hiding and multimedia security* (pp. 5–10).

Bayar, B., & Stamm, M. C. (2018). Constrained convolutional neural networks: A new approach towards general purpose image manipulation detection. *IEEE Transactions on Information Forensics and Security*, *13*(11), 2691–2706.

Bonettini, N., Cannas, E. D., Mandelli, S., Bondi, L., Bestagini, P., & Tubaro, S. (2021). Video face manipulation detection through ensemble of CNNs. In *2020 25th international conference on pattern recognition (ICPR)* (pp. 5012–5019). IEEE.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. In *European conference on computer vision* (pp. 213–229). Springer.

Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A. L., & Zhou, Y. (2021). Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306.

Chen, S., Yao, T., Chen, Y., Ding, S., Li, J., & Ji, R. (2021). Local relation learning for face forgery detection. arXiv preprint arXiv:2105.02577.

Chen, L., Zhang, Y., Song, Y., Liu, L., & Wang, J. (2022). Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 18710–18719).

Chintha, A., Thai, B., Sohrawardi, S. J., Bhatt, K., Hickerson, A., Wright, M., & Ptucha, R. (2020). Recurrent convolutional structures for audio spoof and video deepfake detection. *IEEE Journal of Selected Topics in Signal Processing*, *14*(5), 1024–1037. http://dx.doi.org/10.1109/JSTSP.2020.2999185.

Choi, Y., Choi, M., Kim, M., Ha, J.-W., Kim, S., & Choo, J. (2018). Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 8789–8797).

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1251–1258).

Coccomini, D. A., Messina, N., Gennaro, C., & Falchi, F. (2022). Combining EfficientNet and vision transformers for video deepfake detection. In *International conference on image analysis and processing* (pp. 219–229). Springer.

Cozzolino, D., Poggi, G., & Verdoliva, L. (2017). 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).

Dang, H., Liu, F., Stehouwer, J., Liu, X., & Jain, A. K. (2020). On the detection of digital face manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5781–5790).

Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., & Ferrer, C. C. (2020). The DeepFake detection challenge (DFDC) dataset. arXiv preprint arXiv:2006.07397.

Dolhansky, B., Howes, R., Pflaum, B., Baram, N., & Ferrer, C. C. (2019). The deepfake detection challenge (dfdc) preview dataset. arXiv preprint arXiv:1910.08854.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *International conference on learning representations (ICLR-2021)* (pp. 1–22).

Durall, R., Keuper, M., Pfreundt, F.-J., & Keuper, J. (2019). Unmasking deepfakes with simple features. arXiv preprint arXiv:1911.00686.

Fridrich, J., & Kodovsky, J. (2012). Rich models for steganalysis of digital images. *IEEE Transactions on Information Forensics and Security*, *7*(3), 868–882. http://dx.doi.org/10.1109/TIFS.2012.2190402.

Ganguly, S., Mohiuddin, S., Malakar, S., Cuevas, E., & Sarkar, R. (2022). Visual attention-based deepfake video forgery detection. *Pattern Analysis and Applications*, 1–12.

Güera, D., & Delp, E. J. (2018). Deepfake video detection using recurrent neural networks. In *2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS)* (pp. 1–6). IEEE.

Guo, Z., Yang, G., Chen, J., & Sun, X. (2021). Fake face detection via adaptive manipulation traces extraction network. *Computer Vision and Image Understanding*, *204*, Article 103170.

Heo, Y.-J., Choi, Y.-J., Lee, Y.-W., & Kim, B.-G. (2021). Deepfake detection scheme based on vision transformer and distillation. arXiv preprint arXiv:2104.01353.

Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4401–4410).

Khan, M., Azam, F., Rashid, M., Samea, F., Anwar, M. W., Muzaffar, A. W., & Butt, W. H. (2022). A retargetable model-driven framework for the development of mobile user interfaces. *Journal of Circuits, Systems, and Computers*, *31*(01), Article 2250018.

Koopman, M., Rodriguez, A. M., & Geradts, Z. (2018). Detection of deepfake video manipulation. In *The 20th Irish machine vision and image processing conference (IMVIP)* (pp. 133–136).

Kumar, P., Vatsa, M., & Singh, R. (2020). Detecting face2face facial reenactment in videos. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 2589–2597).

Li, Y., Yang, X., Sun, P., Qi, H., & Lyu, S. (2020). Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3207–3216).

Mohiuddin, S., Ganguly, S., Malakar, S., Kaplun, D., & Sarkar, R. (2022). A feature fusion based deep learning model for deepfake video detection. In *International conference on mathematics and its applications in new computer systems* (pp. 197–206). Springer.

Nguyen, H. H., Yamagishi, J., & Echizen, I. (2019). Use of a capsule network to detect fake images and videos. arXiv preprint arXiv:1910.12467.

Nirkin, Y., Wolf, L., Keller, Y., & Hassner, T. (2021). DeepFake detection based on discrepancies between faces and their context. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1. http://dx.doi.org/10.1109/TPAMI.2021.3093446.

Qian, Y., Yin, G., Sheng, L., Chen, Z., & Shao, J. (2020). Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European conference on computer vision* (pp. 86–103). Springer.

Rahmouni, N., Nozick, V., Yamagishi, J., & Echizen, I. (2017). Distinguishing computer graphics from natural images using convolution neural networks. In *2017 IEEE workshop on information forensics and security (WIFS)* (pp. 1–6). http://dx.doi.org/10.1109/WIFS.2017.8267647.

Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nieß ner, M. (2019). Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 1–11).

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1–9).

Tan, M., & Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning* (pp. 6105–6114). PMLR.

Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *CVPR 2001: vol. 1, Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition* (p. I). IEEE.

Wang, C., & Deng, W. (2021). Representative forgery mining for fake face detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 14923–14932).

Wodajo, D., & Atnafu, S. (2021). Deepfake video detection using convolutional vision transformer. arXiv preprint arXiv:2102.11126.

Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L., & Zhang, L. (2021). Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 22–31).

Yang, X., Li, Y., & Lyu, S. (2019). Exposing deep fakes using inconsistent head poses. In *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 8261–8265). IEEE.

Yousaf, N., Azam, F., Butt, W. H., Anwar, M. W., & Rashid, M. (2019). Automated model-based test case generation for web user interfaces (WUI) from interaction flow modeling language (IFML) models. *IEEE Access*, *7*, 67331–67354.

Yu, C.-M., Chen, K.-C., Chang, C.-T., & Ti, Y.-W. (2022). SegNet: a network for detecting deepfake facial videos. *Multimedia Systems*, *28*, 793—814.

Zhang, K., Zhang, Z., Li, Z., & Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, *23*(10), 1499–1503.