IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 14 November 2023, accepted 5 January 2024, date of publication 11 January 2024,
date of current version 19 January 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3352890

RESEARCH ARTICLE

Improving Video Vision Transformer for Deepfake Video Detection Using Facial Landmark, Depthwise Separable Convolution and Self Attention

KURNIAWAN NUR RAMADHANI[1,2], RINALDI MUNIR[1], AND NUGRAHA PRIYA UTAMA[1], (Member, IEEE)
[1]Bandung Institute of Technology, Bandung 40132, Indonesia
[2]Telkom University, Bandung 40257, Indonesia

Corresponding author: Kurniawan Nur Ramadhani (33219303@mahasiswa.itb.ac.id)

ABSTRACT In this paper, we present our result of research in video deepfake detection. We built a deepfake detection system to detect whether a video is a deepfake or real. The deepfake detection algorithm still struggle in providing a sufficient accuracy values, especially in challenging deepfake dataset. Our deepfake detection system utilized spatiotemporal feature that extracted using Video Vision Transformer (ViViT). The main contribution of our research is providing a deepfake detection system that based on ViViT architecture and using landmark area images for the input of the system. Our system extracted the feature from a number of spatial features. The spatial feature was extracted using Depthwise Separable Convolution (DSC) block combined with Convolution Block Attention Module (CBAM) from tubelet. The tubelet was a representation of facial landmark area that was extracted from the input video. In our system, we used 25 facial landmark area for an input video. In our experiment we used Celeb-DF version 2 dataset because it is considered to be a challenging deepfake dataset. We conducted augmentation to the dataset, so we obtained 8335 videos for training set, 390 videos for validation set, and 1123 videos for testing set. We trained our deepfake detection system using Adam optimizer, with learning rate of 10-4 and 100 epoch. From the experiment, we obtained the accuracy score of 87.18% and F1 score of 92.52%. We also conducted the ablation study to display the effect of each part of our model to the overall system performance. From this research, we obtained that by using landmark area images, our ViViT based deepfake detection system had a good performance in detecting deepfake videos.

INDEX TERMS Deepfake detection, facial landmark, depthwise separable convolution, convolution block attention module, video vision transformer.

I. INTRODUCTION

The development of algorithms in the field of machine learning in various fields has allowed humans to build applications that can do things that were previously difficult for existing computers to do. Particularly in the field of computer vision, the development of machine learning algorithms has made it possible for technology to manipulate and even create images and videos at a further level. This is in addition to providing a positive effect also has a negative impact. Deepfake is one of the negative effects that arise from the development of computer vision technology.

Deepfakes are content in the form of images or videos that are manipulated using deep learning algorithms. Usually deepfake content manipulates videos of people doing something. The manipulation process that is carried out is by exchanging the faces of people who are in the video with other people's faces. In its implementation, deepfake can be used for several purposes. Several deepfake videos circulating on social networking applications feature certain clips of films

The associate editor coordinating the review of this manuscript and approving it for publication was Andrea Bottino.

with the faces of actors replaced with the faces of other people. Some of these videos are just for fun. However, some deepfake videos have quite serious negative effects.

Two cases that often become objects of deepfake are pornographic videos and black campaign videos. In the case of pornographic videos, deepfakes manipulate pornographic videos by replacing the actor's face with another face, which is usually the face of a particular artist or public figure. The goal is to drop the image of the character. Another case is a political campaign video. In the case of black campaign videos, deepfakes manipulate videos of someone making controversial statements by replacing the speaker's face with another face, which is usually the face of a certain political figure who is participating in a political contest. The aim is to drop the electability of certain political figures. These two cases are examples of the harmful use of deepfakes.

Looking at these two cases, technology to detect deepfake content in images and videos is indispensable. Deepfake detection technology will be very useful to protect people if they fall victim to the wrong use of deepfake technology. Since 2017, many studies have been conducted to develop deepfake detection methods, both in images and in videos. The deepfake detection method basically attempts to classify image or video content into the fake class or the original class. The deepfake detection algorithm attempts to extract features from image or video content that can be used to distinguish fake content from original content.

In general, there are four feature extraction approaches that have been developed to detect deepfakes. The first approach is to use visual traits, which are traits that appear directly from deepfake content. These visual characteristics can be in the form of eye blinks, head position and other characteristics that can be observed on the face. The deepfake detection approach with an eye wink gets an AUC score of 0.99. The weakness of this approach is that the detection process is highly dependent on the eye area location detection algorithm and the test data used is only 10 videos [1]. The deepfake detection approach using head pose information managed to get an AUC score of 0.89 in the UADFV dataset. Another visual artifact approach succeeded in achieving an AUC score of 0.85 on deepfake test data derived from 4 original videos [2].

The second approach is to use local features, namely features extracted by certain methods at the image pixel level. Local feature extraction method has better robustness than visual features. Research conducted in 2019 compared several local feature extraction approaches in detecting deepfakes and concluded that Image Quality Metric (IQM) is the best local feature extraction for detecting deepfakes with an Equal Error Rate (EER) score of 8.97% in the DF-TIMIT [3].

The third approach is to use deep features. As with local feature extraction, deep feature extracts features at the pixel level, but uses deep learning algorithms so that the resulting features are more complex. Several deep feature models that have been developed to detect deepfakes include MesoNet [4], DeepFD [5], Common Fake Feature Network (CFFN) [6], MultiTask Learning [7] and Capsule Forensic [8].

The fourth approach is to use a temporal feature. This approach only works with videos. This approach, in addition to using features at the pixel level, also uses temporal information from several successive frames. Examples of this approach include using a Recurrent Neural Network (RNN) [9] and Optical Flow [10].

The main challenge of deepfake detection algorithm development is the ever-evolving deepfake content development algorithm. With the development of the deepfake algorithm, the ability of the deepfake detection method must continue to be improved. This can be observed from the development of the dataset used in deepfake research. In 2019, the Celeb-DF dataset was used to test several deepfake algorithms [11]. The results of the test show that the deepfake detection algorithms that have been developed have not succeeded in showing good performance in detecting deepfakes on the Celeb-DF dataset with an accuracy value of all algorithms tested less than 70%. This low accuracy value is caused by the generalization ability of each deepfake detection method which is still weak. In this research, the deepfake detection algorithm that has the highest accuracy is the XceptionNet [12] model which is a deep feature approach.

In this research, we built a system to detect deepfake videos using spatiotemporal approach. We used Video Vision Transformer (ViViT) [13] to detect the deepfake videos. We used 25 facial landmark location as input for our ViViT model. In the encoding part of our ViViT, we used Depthwise Separable Convolution (DSC) [12] block combined with Convolution Block Attention Module (CBAM) [14] to extract feature from the tubelet. We used Celeb-DF version 2 for the dataset of our research. In brief, our contribution in this paper as follows:

a. We built deepfake video detection system using Video Vision Transformer (ViViT)
b. We used 25 facial landmark extracted from dataset as our input for the detection system
c. We used DSC combined with CBAM to extract feature from the tubelet before the positional encoding block in ViViT

II. RELATED WORKS

The detection of deepfake content is a binary process that involves the extraction of features from images and videos to differentiate between genuine and deepfake content. This method can be classified into four distinct categories, depending on the particular feature extraction approach employed.

A. VISUAL FEATURE-BASED DEEPFAKE DETECTION

This approach relies on facial feature that can be observed in plain sight, such as head pose, eye blink, and dissemblance in facial organ shape. Research using this approach were first conducted in 2018 that used eye blink for the main feature [1]. The hypothesis behind the method is that there are variations in the blink pattern of the deepfake compared to the original video.

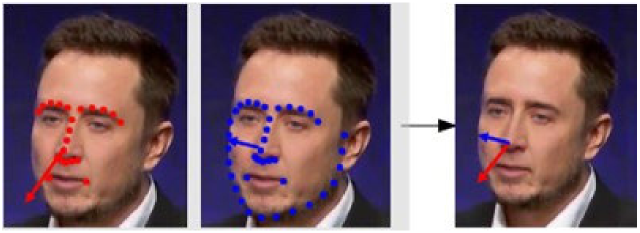

FIGURE 1. **Head pose inconsistency estimation.**

Another research used head pose inconsistencies to detect deepfakes [15]. This approach considers inconsistencies between facial poses as well as body parts outside of the face such as neck and shoulders. Figure 1 shows how to use 68 facial landmarks to estimate head pose inconsistency. The 68 facial landmarks (68 blue landmarks) were compared to 17 red landmarks (17 red landmarks) which represent the pose direction from the midpoint facial area.

Another approach tries to extract visual features on the deepfake face [2]. These were referred to as visual artifacts. Visual artifacts are extracted from the imperfectness of the deepfake because the resources were limited in the process of creating the deepfake. Visual artifacts include left eye color difference, right eye color difference, disproportionate shadows, fack of light reflection detail, and lack of detailed geometry. Visual artifacts such as left eye color difference, right eye color difference, disproportionate shadows in nose area, non-visible light reflections, and less detailed geometry of teeth. Visual artifacts were extracted using a color and geometry extraction approach from precise parts of the face, such as nose, lips, teeth, eyes, and eyebrows.

It is possible to use this visual feature approach to detect a deepfake. However, as the content creation methods for deepfakes become more advanced, the detection of these visual features becomes more difficult. As a result, these deepfakes detection methods based on visual features become less effective.

B. LOCAL FEATURE-BASED DEEPFAKE DETECTION

This approach used pixel-based segmentation to extract features from each pixel of an image. Local feature-based detection is more reliable than visual features. In 2017, a research combined two features from an image convolution and two features from image steganalysis to identify tampered areas in a facial image [16]. This research formed the fundament for local and deep feature deepfake detection methods.

Another approach was to use photo response non uniformity analysis with cross correlation operations (PRNU) [17]. However, this research used a dataset of only 10 videos. Another feature extraction method used to detect deepfakes is the scale-invariant feature transform (SIFT) [18]. SIFT identifies pixel keypoints in an image and extracts features from those keypoints.

Other feature extraction methods used in the deepfakes detection process include Pyramid of Histogram of Oriented Gradients (PHOG), Local Phase Quantization (LPQ), Local Binary Pattern (LBP), Speeded Up Robust Feature (SURF), Binary Gabor pattern (BGP), Binarized Statistical Image Features (BSIF), and Image Quality Metric (IQM) [19]. The IQM method is the most effective way to detect deepfakes, and this was confirmed in a research that compare IQM's to LDA's and PCA's, in which IQM was the most effective feature.

The local feature-based detection method has been found to be quite effective in detecting the presence of deepfakes in certain video data. However, with the development of deepfakes algorithms, the tampered contents tend to be seen more natural and hard to detect as a deepfake. More complex features were required to differentiate the original image and video.

C. DEEP FEATURE-BASED DEEPFAKE DETECTION

Similar to local features, deep features also performs pixel-level feature extraction. The difference is that the deep feature extraction process uses many layers, meaning that it capable to obtain more complex features than simple feature extraction techniques. In 2018, a research analyzed DenseNet, InceptionNet, and XceptionNet to detect deepfake images [20]. The XceptionNet architecture was found to be the most reliable in detecting deepfakes. Another research used CNN architecture with 5 layers, named DeepFD, with a contrastive loss function providing good performance for detecting various GAN generated images [5]. Further research expanded DeepFD with a pair learning method that increased the generalization ability of the model [6]. In a pair learning approach, two pairs of real or fake images are analyzed using contrastive learning. The new model were called Common Feature Fake Network (CFFN).

MesoNet was another CNN model based on the inception module for detecting deepfakes [4]. This model was able to detect deepfake videos with compression conditions similar to social media videos. However, research using Capsule Network architecture can compete with MesoNet for detecting deepfake videos at per frame level and whole video level [8]. Capsule Network compensates for non-equivocation convolution blocks by routing by agreement mechanism.

Another deep learning model that has been developed for the detection of deepfake is deep autoencoder [7]. The autoencoder was able to reconstruct contents and mark the deepfake areas in the contents. In this research, the Autoencoder used two branches. One branch was used to reconstruct the deepfake marking. The other was used to compute the loss function of the Autocoder. Segmentation of the image area improves the detection result compared to just detecting the deepfake or the original content.

D. TEMPORAL FEATURE-BASED DEEPFAKE DETECTION

Unlike the other approaches, this method extracts features from multiple sequential frames to get temporal features.

Thus, this method is applicable specifically to video. Temporal feature approach works by extracting temporal features from a sequence of video frames. A sequence of video frames can be considered as a sequence of data.

Recurrent Neural Network (RNN) is one of the most well-known deep neural network (DNN) models for processing sequence data. RNN was first used to detect deepfake content in videos in 2018 [9]. Another method of deepfake video detection is to track facial and head movement and then extract 16 motion features from a specific area in the video [21]. In addition to RNN and motion tracking, Optical Flow method has been used in deepfake video detection research with CNN as the classification algorithm [10]. RNN then was re-used in 2019 by using two-way RNN and DenseNet which outperforms several DNN models (ResNet50, DenseNet) [22].

III. METHODS

In this research, we built our deepfake detection system based on Video Vision Transformer (ViViT) architecture [13]. The overall system is shown in Figure 2.

The input of our system is video. Our method consist of several steps. The first step is the preprocessing, where the system extracts the 25 facial landmark area from the video and combines them into sequence of frames. Each frame contains the 25 facial landmark area for respective frame from the input video. Next, the system build tubelet from the frame sequence. For each tubelet, the system extract the features using Depthwise Separable Convolution (DSC) block combined with Convolution Based Attention Module (CBAM). The features then flattened and encoded using positional encoder, then processed by spatiotemporal analysis in the core of ViViT. The core itself contain of several Multihead Attention Block. The final process is the classification using sigmoid function.

A. FACIAL LANDMARK DETECTION

For this purpose, we used the facial landmark detection operation in dlib library [23]. Facial landmark detection operation is used to obtain a total of 68 landmark points on facial images which contain important information, especially facial curves, both on the edges of the face and areas inside the face, such as the curves of the nose, lips and eyes. The facial landmark detection process in Dlib uses the Histogram of Oriented Gradients (HOG) algorithm and also the Support Vector Machine (SVM) classification. In this research, we used only 25 landmark to simplify the input data for the system. For each landmark, we extracted 11 × 11 area centered by landmark location pixel. So, for one frame, we extracted new 55 × 55 frame that contain the information of each 11 × 11 area from 25 landmark location. The landmark location that we used can be seen on Figure 3.

The landmark area only consider the area that marked by the landmark point. By that consideration, the landmark area images lost some information from the original input. Our model only conserve the tubelet from landmark area. So, patch position information produced by the positional encoder was only contained the position information of landmark area. The tubelet sequence in our model did not preserve the original input's patch position. But, by using landmark area images, the ViViT was not overbloated by unnecessary information from the non landmark area, as we proved this statement by the ablation study in the section V.

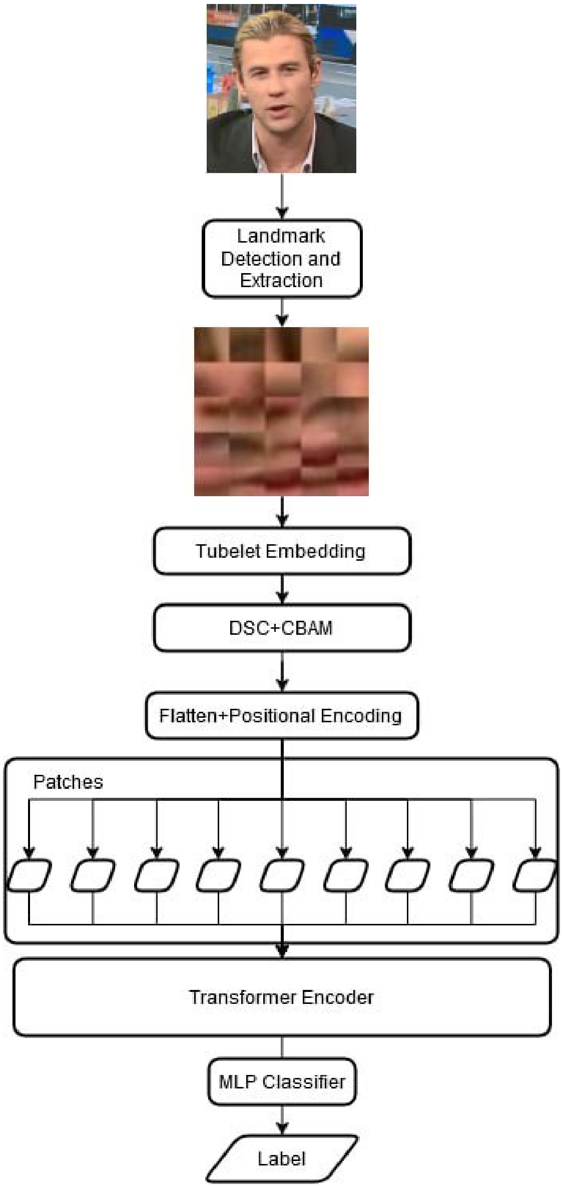

FIGURE 2. Our deepfake detection system.

B. DEPTHWISE SEPARABLE CONVOLUTION (DSC)

DSC is a convolution block module used in the Xception model that substitute the function of Inception module [12]. Depthwise separable convolution has 1 × 1 depthwise convolution and pointwise convolution. Depthwise convolution is a separate convolution process for each input channel.

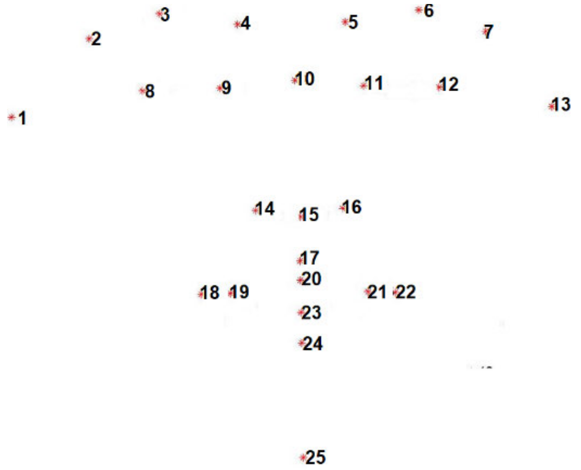

FIGURE 3. Facial landmark location.

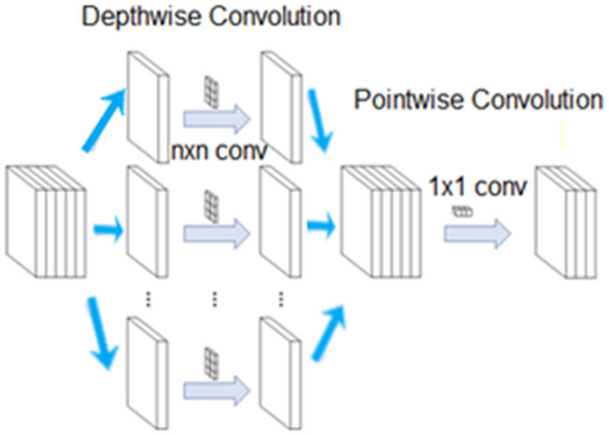

FIGURE 4. Depthwise separable convolution.

Pointwise convolution is a convolution using 1 × 1 block on all input channel. The detail of DSC is described in Figure 4.

The first step in this DSC is to perform convolution on each channel in turn. Then in the next step is to perform pointwise convolution which was a standard convolution with the 1 × 1 kernel. Using depthwise separable convolution can reduce the number of mathematical operations and the number of parameters used in the process.

C. CONVOLUTION BLOCK ATTENTION MODULE (CBAM)

In this research, beside using DSC, we also used CBAM model to extract feature from tubelet. This module is a simple attention model specifically designed for the CNN architecture. CBAM consists of two processes, namely the channel attention module and the spatial attention module [14]. Figure 5 shows the CBAM architecture.

Channel Attention Module is a series of operations to generate channel attention maps. The channel attention map represents the strength of the relationship between channels from the input data features. The channel attention module begins with max pool and average pool operations to get a representation of the value of each channel. Then the two vectors are processed using the same neural network then the results of the two are added up and multiplied by the initial features to get the channel attention map.

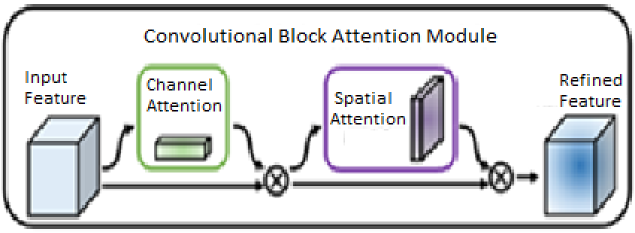

FIGURE 5. CBAM architecture.

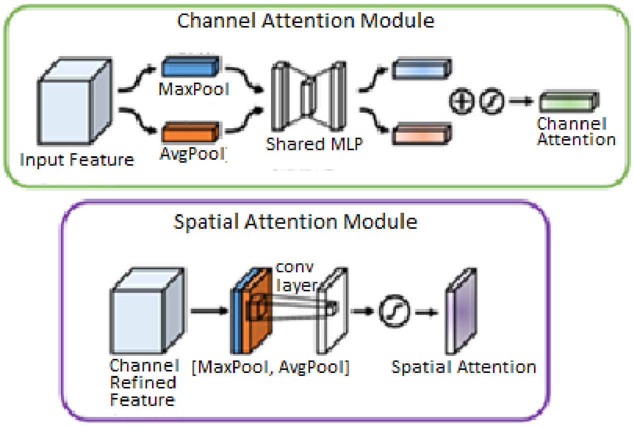

FIGURE 6. Channel and spatial attention module.

The spatial attention module is used to obtain spatial attention map values. In CBAM, the spatial attention module processes values from the channel attention map generated from the spatial attention module. The spatial attention module begins with the max pool and average pool operations. The two pooling results are connected and then convoluted into one feature map. The feature map is then multiplied by the input features to produce a spatial attention map. Figure 6 shows the channel and spatial attention module.

While the ViViT already provide the self-attention mechanism, the attention value provided by ViViT is only the attention value between different encoded tubelet. The using of CBAM was to provide the attention value between the feature values in one tubelet. As has been mentioned before, CBAM provide the attention for channel and spatial dimension. So in our model, the value of one tubelet was enriched by the spatial and channel attention value provided by CBAM.

D. VIDEO VISION TRANSFORMER

Vision Transformer (ViT) receives sequential data in image fragments for classification using the transformer encoder module, which is then decoded by the linear embedding layer belonging to the transformer model. Figure 7 shows the general form of the ViT model with layers of embedding, encoding, and classifier. Different from CNN model in which

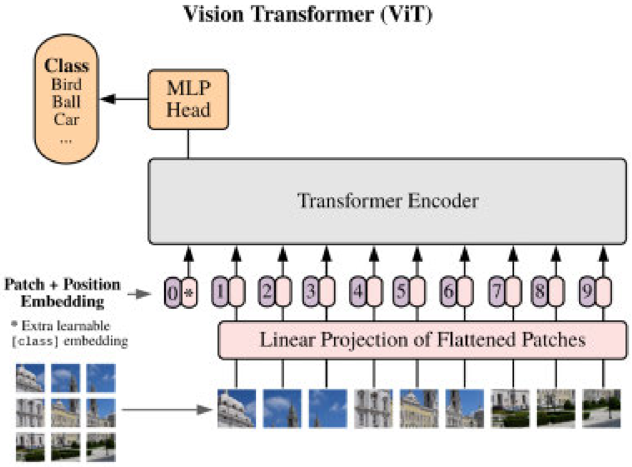
Vision Transformer (ViT)
Class
Bird
Ball
Car
...
MLP Head
Transformer Encoder
Patch + Position Embedding
* Extra learnable [class] embedding
0*
1
2
3
4
5
6
7
8
9
Linear Projection of Flattened Patches

FIGURE 7. Conventional vision transformer.

uses spatial space with a kernel and local receptive field, ViT can use attention in a part of the image that is broken down by converting the image data into vectors, because this ViT requires large datasets to obtain sufficient spatial knowledge [24].

The first step is to partition the training dataset into patches. Every patches then flattened to become token. Because our data is in video form, we used tubelet for representation of patch. We extracted the volume from input video that contain image patches as well as temporal information from the video. In our system, every tubelet represented the information of respective landmark location, so for one input video we extracted 25 tubelet.

IV. EXPERIMENTAL SETUP

In this research, we used the Celeb-DF dataset version 2 [11]. We chose this dataset because of its variation and challenging characteristics in deepfake detection research. In the dataset, there are 890 original videos and 5639 deepfake videos. We processed the dataset using the facial landmark detection function from dlib to obtain 25 landmark locations. From the landmark locations, we took area of 11 × 11 with the landmark location as the center of the area, then we combined the 25 areas into one new frame with a size of 55 × 55. From one video, we took 55 frames so that the dimensions of the input video on our system were 55 × 55 × 55 × 3, where 3 is the RGB channel. For the training set, we carried out an augmentation process in the form of horizontal flip, random rotation and increasing the pixel value so that the total training set we used was 8335 videos. We also used 390 videos for validation set in the training process. For the testing set, we took 1123 videos with a proportion of 139 original videos and 984 deepfake videos. Figure 8 shows the example of original face and it's new frame created from our process.

The development and testing environment that we used in this research was Google Colaboratory pro. Google Colaboratory pro has Nvidia A100 GPU with 80GB of VRAM and 32GB of RAM. The parameters that we used in the learning process were learning rate of 10^{-4}, batch size of 8, epoch of 100 and Adam as our optimizer algorithm.

Other hyper parameters that we used in the network architecture were patch size of 11 × 11×11, 128 projection dimensions, 32 attention heads and 16 attention layers.

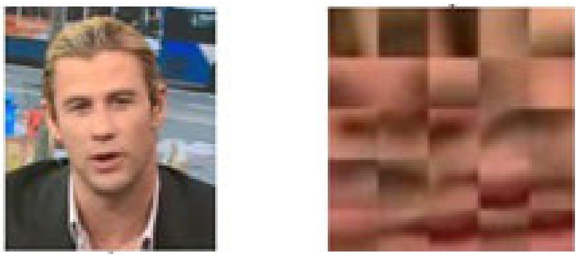

FIGURE 8. Example of original face and landmark area extraction result.

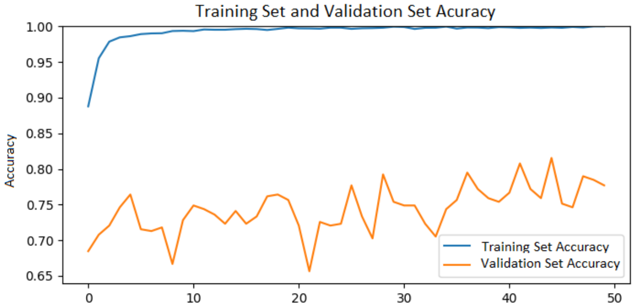

FIGURE 9. Training set and validation set accuracy over the training process.

V. RESULT AND DISCUSSION

In this section, we present the result of our experiment. Figure 9 shows the validation and training accuracy over the training process. We can see that over the training process, the training accuracy was convergent to 100% and the validation accuracy was unstable between 70% and 80%. We can see that our system still had some overfitting issue, although it's not significant.

Next we can see the performance of our system in the testing process. Figure 10 shows the confusion matrix for our system. We can see that our system successfully recognized 890 deepfake videos from 984 deepfake videos in the testing set. Our system also recognized 89 real videos from 139 real videos in the testing set. We can see that our system had a good performance in the testing process, with F1 score of 92.52% and accuracy of 87.18%.

In our research, we compared our system with conventional ViViT without using DSC+CBAM. We can see the comparison of accuracy in table 1. As can be seen, our system had better performance compared to conventional ViViT. The DSC+CBAM gave positive effect on detection system performance. By combining DSC and CBAM, we obtained a good feature extractor for each landmark area tubelet.

We also compared our system with other deepfake detection system. We compared our system with Mesonet [4],

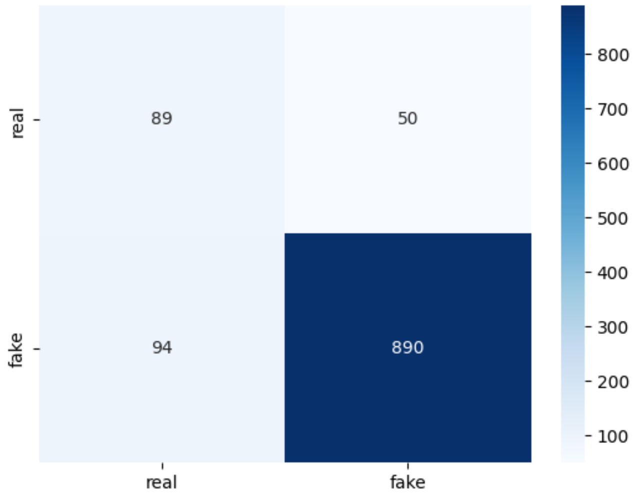

FIGURE 10. Confusion matrix of our deepfake detection system.

TABLE 1. Comparison between our system and ViViT.

Detector	Accuracy
ViViT	52.14%
Our system	**87.18%**

TABLE 2. Comparison between our system and other deepfake detection system.

Detector	Accuracy
Mesonet	65.12%
Xception	72.28%
Xception+CBAM	75.49%
Our system	**87.18%**

Xception [12], and Xception+CBAM [25]. The comparison result can be seen on table 2. From the table, we can see that our system had better performance in term of accuracy. This result showed that the spatiotemporal features that extracted from our model can improve the performance of deepfake detection system. The use of landmark area also increased the performance by narrowing the extraction area in the feature extraction process.

We also conducted ablation study in our research. We compared our method with the other combination of ViT, landmark extraction, DSC and CBAM as follows:

a. version 1 that used ViT, DSC and CBAM module.
b. Version 2 that used ViT, landmark extraction and DSC module.
c. Version 3 that used ViT and landmark extraction.
d. Version 4 that used ViT and DSC module.
e. Version 5 that used only ViT.

Table 3 showed the comparison of result from the ablation study. By comparing our method result and version 2 result, and also comparing result of version 1 and version 4, we concluded that CBAM module gave a positive effect on the accuracy and F1-score. By comparing version 2 and version 3 result, we concluded that DSC module also gave a positive effect on the system performance. By comparing the result of our method and version 1, and also by comparing result of version 2 and version 4, and also the result of version 3 and version 5, we concluded that the landmark area extracted from input images had significant information that boosted the performance of ViViT for detecting deepfake.

TABLE 3. Ablation study result.

Model	ViViT	landmark extraction	DSC	CBAM	acc	F1 score
Version 1	√		√	√	73.12%	80.67%
Version 2	√	√	√		75.71%	84.01%
Version 3	√	√			68.76%	79.63%
Version 4	√		√		67.74%	74.84%
Version 5	√				52.14%	63.60%
our method	**√**	**√**	**√**	**√**	**87.17%**	**92.51%**

TABLE 4. Performance comparison for several different dataset.

Dataset	Accuracy	F1 score
Celeb-DF version 2	87.17%	92.51%
Deepfake Detection Challenge (DFDC)	88.03%	90.23%
Deepfake TIMIT	94.14%	97.39%
FaceForensics++	90.27%	93.54%

Apart from Celeb-DF version 2, we also evaluated our model using several different public dataset such as Deepfake Detection Challenge (DFDC) dataset [26], Deepfake TIMIT dataset [27], and FaceForensics++ deepfake dataset [28]. Table 4 displayed the performance of our system over those dataset. We can see that our model had a consistent performance in terms of accuracy and F1 score over different deepfake dataset.

VI. CONCLUSION

In this research, we built a deepfake detection system that detect whether the video input is a deepfake video or real video. We extracted 25 landmark area from the input videos before processed into our deepfake detection system. We combined DSC and CBAM to extract the spatial features from each of landmark area. Then, the spatial features were processed using ViViT to detect deepfake. From the experiment, our system shown a good result by obtaining accuracy of 87.18%. Our system successfully detected 890 from 984 deepfake videos and 89 from 139 real videos. Our system also successfully outperform some other deepfake detection system, such as Mesonet, Xception and Xception+CBAM. Our approach by extracting 25 landmark area from videos has been proven to improve the performance of deepfake detection system by reducing unimportant area for feature extraction process. The combination of Xception and CBAM was also proven to be a suitable feature extractor for the landmark area tubelet. The spatiotemporal extracted from our ViViT also improved the performance in detecting deepfake videos.

REFERENCES

[1] Y. Li, M.-C. Chang, and S. Lyu, ''In ictu oculi: Exposing AI created fake videos by detecting eye blinking,'' in *Proc. IEEE Int. Workshop Inf. Forensics Secur. (WIFS)*. Hong Kong: IEEE, Dec. 2018, pp. 1–7, doi: 10.1109/WIFS.2018.8630787.

[2] F. Matern, C. Riess, and M. Stamminger, ''Exploiting visual artifacts to expose deepfakes and face manipulations,'' in *Proc. IEEE Winter Appl. Comput. Vis. Workshops (WACVW)*. Waikoloa Village, HI, USA: IEEE, Jan. 2019, pp. 83–92, doi: 10.1109/WACVW.2019.00020.

[3] Z. Akhtar and D. Dasgupta, ''A comparative evaluation of local feature descriptors for DeepFakes detection,'' in *Proc. IEEE Int. Symp. Technol. for Homeland Secur. (HST)*. Woburn, MA, USA: IEEE, Nov. 2019, pp. 1–5, doi: 10.1109/HST47167.2019.9033005.

[4] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, ''MesoNet: A compact facial video forgery detection network,'' in *Proc. IEEE Int. Workshop Inf. Forensics Secur. (WIFS)*. Hong Kong: IEEE, Dec. 2018, pp. 1–7, doi: 10.1109/WIFS.2018.8630761.

[5] C.-C. Hsu, C.-Y. Lee, and Y.-X. Zhuang, ''Learning to detect fake face images in the wild,'' in *Proc. Int. Symp. Comput., Consum. Control (IS3C)*. Taichung, Taiwan: IEEE, Dec. 2018, pp. 388–391, doi: 10.1109/IS3C.2018.00104.

[6] C.-C. Hsu, Y.-X. Zhuang, and C.-Y. Lee, ''Deep fake image detection based on pairwise learning,'' *Appl. Sci.*, vol. 10, no. 1, p. 370, Jan. 2020, doi: 10.3390/app10010370.

[7] H. H. Nguyen, F. Fang, J. Yamagishi, and I. Echizen, ''Multi-task learning for detecting and segmenting manipulated facial images and videos,'' in *Proc. IEEE Int. Conf. Biometrics, Theory, Appl. Syst. (BTAS)*, Jun. 2019, pp. 1–8. Accessed: Apr. 1, 2020. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9185974/

[8] H. H. Nguyen, J. Yamagishi, and I. Echizen, ''Capsule-forensics: Using capsule networks to detect forged images and videos,'' in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*. Brighton, U.K.: IEEE, May 2019, pp. 2307–2311, doi: 10.1109/ICASSP.2019.8682602.

[9] D. Güera and E. J. Delp, ''Deepfake video detection using recurrent neural networks,'' in *Proc. 15th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*. Auckland, New Zealand: IEEE, Nov. 2018, pp. 1–6, doi: 10.1109/AVSS.2018.8639163.

[10] I. Amerini, L. Galteri, R. Caldelli, and A. D. Bimbo, ''Deepfake video detection through optical flow based CNN,'' in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1205–1207, doi: 10.1109/ICCVW.2019.00152.

[11] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, ''Celeb-DF: A large-scale challenging dataset for DeepFake forensics,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3204–3213, doi: 10.1109/CVPR42600.2020.00327.

[12] F. Chollet, ''Xception: Deep learning with depthwise separable convolutions,'' 2016, *arXiv:1610.02357*.

[13] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, ''ViViT: A video vision transformer,'' 2021, arXiv:2103.15691.

[14] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, ''CBAM: Convolutional block attention module,'' 2018, *arXiv:1807.06521*.

[15] X. Yang, Y. Li, and S. Lyu, ''Exposing deep fakes using inconsistent head poses,'' in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*. Brighton, U.K.: IEEE, May 2019, pp. 8261–8265, doi: 10.1109/ICASSP.2019.8683164.

[16] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis, ''Two-stream neural networks for tampered face detection,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Mar. 2018, pp. 1831–1839. Accessed: Apr. 1, 2020. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8014963/

[17] M. Koopman, A. M. Rodriguez, and Z. Geradts, ''Detection of deepfake video manipulation,'' in *Proc. 20th Irish Mach. Vis. Image Process. Conf. (IMVIP)*, 2018, p. 4.

[18] G. Wang, Q. Jiang, X. Jin, and X. Cui, ''FFR_FD: Effective and fast detection of DeepFakes via feature point defects,'' *Inf. Sci.*, vol. 596, pp. 472–488, Jun. 2022, doi: 10.1016/j.ins.2022.03.026.

[19] P. Korshunov and S. Marcel, ''Vulnerability assessment and detection of deepfake videos,'' in *Proc. Int. Conf. Biometrics (ICB)*. Crete, Greece: IEEE, Jun. 2019, pp. 1–6, doi: 10.1109/ICB45273.2019.8987375.

[20] F. Marra, D. Gragnaniello, D. Cozzolino, and L. Verdoliva, ''Detection of GAN-generated fake images over social networks,'' in *Proc. IEEE Conf. Multimedia Inf. Process. Retr. (MIPR)*. Miami, FL, USA: IEEE, Apr. 2018, pp. 384–389, doi: 10.1109/MIPR.2018.00084.

[21] S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano, and H. Li, ''Protecting world leaders against deep fakes,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2019, pp. 38–45.

[22] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, and P. Natarajan, ''Recurrent convolutional strategies for face manipulation detection in videos,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2019, pp. 80–87.

[23] D. Zhang, J. Li, and Z. Shan, ''Implementation of Dlib deep learning face recognition technology,'' in *Proc. Int. Conf. Robots Intell. Syst. (ICRIS)*, Nov. 2020, pp. 88–91, doi: 10.1109/ICRIS52159.2020.00030.

[24] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, Z. Yang, Y. Zhang, and D. Tao, ''A survey on vision transformer,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 87–110, Jan. 2023, doi: 10.1109/TPAMI.2022.3152247.

[25] H. Lin, W. Luo, K. Wei, and M. Liu, ''Improved xception with dual attention mechanism and feature fusion for face forgery detection,'' in *Proc. 4th Int. Conf. Data Intell. Secur. (ICDIS)*, Aug. 2022, pp. 208–212, doi: 10.1109/ICDIS55630.2022.00039.

[26] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. C. Ferrer, ''The DeepFake detection challenge (DFDC) dataset,'' 2020, *arXiv:2006.07397*.

[27] P. Korshunov and S. Marcel, ''DeepFakes: A new threat to face recognition? Assessment and detection,'' 2018, *arXiv:1812.08685*.

[28] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, ''FaceForensics++: Learning to detect manipulated facial images,'' in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1–11.

KURNIAWAN NUR RAMADHANI is currently pursuing the Ph.D. degree with the School of Electrical Engineering and Informatics, Bandung Institute of Technology. His research interests include computer vision and machine learning.

RINALDI MUNIR received the bachelor's degree in informatics engineering and the M.Sc. degree in digital image compression from the Bandung Institute of Technology (ITB), Bandung, Indonesia, in 1992 and 1999, respectively, and the Ph.D. degree in image watermarking from the School of Electrical Engineering and Informatics, ITB, in 2010. In 1993, he started his academic career as a Lecturer with the Department of Informatics, ITB. He is currently an Associate Professor with the School of Electrical Engineering and Informatics, ITB, and the Informatics Research Group. His research interests include cryptography and steganography-related topics, digital image processing, fuzzy logic, and numerical computation.

NUGRAHA PRIYA UTAMA (Member, IEEE) received the bachelor's degree in informatics from the Bandung Institute of Technology, Indonesia, in 2002, and the master's and Ph.D. degrees from the Tokyo Institute of Technology, in 2006 and 2009, respectively. His research interests include computer vision and neuroscience.