Received: May 29, 2025. Revised: July 20, 2025.

INASS

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

VERITAS: Vision-based Excitation and Robust Intelligence for Transformer-Assisted Deepfakes Detection

Alam Rahmatulloh[1,2] Herman Dwi Surjono[1] Fatchul Arifin[1] Irfan Darmawan[3]* Nia Ambarsari[3]

[1]*Faculty of Engineering, Universitas Negeri Yogyakarta, Yogyakarta, Indonesia*
[2]*Faculty of Engineering, Universitas Siliwangi, Tasikmalaya, Indonesia*
[3]*Faculty of Industrial Engineering, Telkom University, Bandung, Indonesia*
* Corresponding author's Email: irfandarmawan@telkomuniversity.ac.id

Abstract: The increasingly massive proliferation of deepfake content poses a serious threat to the authenticity and trustworthiness of digital information. This study proposes VERITAS, a deepfake detection framework that integrates CNN-based feature extraction, Vision Transformer (ViT) architecture, and multilevel Squeeze-and-Excitation (SE) blocks to adaptively strengthen spatial and token attention. To enhance generalization across datasets, VERITAS also adopts a dual-branch self-learning mechanism consisting of Masked Image Modeling (MIM) and Identity Distillation (ID) based on the Face Security Foundation Model (FSFM) framework. Experimental results show that VERITAS achieves AUC scores of 87.45% and 88.30% on the Celeb-DF v2 and WildDeepfake datasets at frame-level, and 95.56% and 87.75% at video-level, respectively, outperforming various state-of-the-art methods. Ablation studies confirm the significant contributions of each component of the architecture. Despite its strong detection performance, inference speed remains a challenge, so further research directions include optimizing the model for real-time deployment. This research contributes to building a robust and adaptive deepfake detection system across domains.

Keywords: CNN, Deepfake detection, FSFM framework, Squeeze-and-excitation, Vision transformer.

1. Introduction

Deepfake technology powered by generative adversarial networks (GANs) and deep learning models has revolutionized content creation by enabling the creation of highly realistic synthetic media [1]. However, the misuse of this technology for the production of misleading content, political propaganda, and identity fraud presents serious challenges. To overcome these negative impacts, a reliable deepfake detection system is needed [2]. Despite efforts to develop various solutions, there are still some major open issues related to hidden deepfakes. CNN-based models tend to focus too much on a particular form of falsification, resulting in poor performance when tested on different datasets [3]. Furthermore, most current detection models remain vulnerable to, for example, video compression and noise addition, which, from the perspective of model performance in real life, make them very poor in operation [4]. Another difficult problem is how to represent features. Convolutional Neural Networks (CNN) are known to be very good at capturing local spatial features, but very poor at capturing global spatial relationships needed to detect subtle discrepancies that are characteristic of deepfake videos [5].

Vision Transformer (ViT) has attracted much interest due to its ability to capture global dependencies of an image, unlike CNNs that are more oriented towards local features [5]. Despite these advances, one of the critical drawbacks of ViT remains its reliance on large datasets to effectively train the model and avoid overfitting. Thus, optimization strategies are needed to make ViT-based models more efficient in feature extraction on limited datasets. Several researchers have applied transformer-based approaches to deepfake detection,

such as Hybrid Transformer Network [6] that utilizes CNN as a feature extractor before the data is transformed by the transformer. In addition, Transformer-Based Feature Compensation and Aggregation [7] designs blocks to compensate for the local feature scope so that the model can better detect subtle fraud patterns. DSViT: An Enhanced Transformer Model [8] also proposes SC-Convolution with ViT to improve detection accuracy. Hybrid approaches combining CNN and Transformer have been shown to improve the detection accuracy of deepfake videos [9].

In previous studies, we found that CNN-based models are quite good at extracting local features, but not so good in the context of global information. On the other hand, transformer models such as ViT are considered superior in capturing further spatial dependencies, but are highly dependent on the quantity of data and available computational capabilities [10]. Several studies have also indicated that CNN-based or ViT-based approaches individually are not good enough in dealing with increasingly complex deepfakes [11].

However, these studies have several limitations, such as the need for large datasets, sensitivity to noise, and lack of adaptive techniques for different types of deepfakes. To address some of the major shortcomings of previous studies, this study proposes a hybrid CNN-Transformer deepfake detection model with Squeeze-and-Excitation Attention (SE-Blocks). With SE-Blocks, the model becomes more sensitive to relevant features, unlike previous efforts where large datasets are required for ViT to learn effectively. As cited in [12], this study also uses self-supervised contrastive learning and adversarial training to improve model generalization and reduce vulnerability to adversarial manipulation. It is expected that with this approach, the proposed model will be able to improve the effectiveness of deepfake detection compared to current methods in generalizing across datasets and in mitigating adversarial attacks. Therefore, this research makes a significant contribution towards the development of a more reliable deepfake detection system that can function under a variety of real-world conditions.

2. Related works

In recent years, deepfakes have developed significantly, allowing for face and voice manipulation that is almost undetectable by humans. This development raises serious challenges in digital security, privacy, and disinformation [13]. In an effort to address this, various detection approaches have been proposed, ranging from conventional visual feature-based methods to deep learning models such as CNN [14]. In addition, Vision-Transformer (ViT)-based approaches are gaining more attention due to their ability to capture spatial dependencies more broadly [15]. This section discusses in detail the latest CNN and transformer-based approaches that are the foundation for the development of the VERITAS (Vision-based Excitation and Robust Intelligence for Transformer-Assisted Deepfakes Detection) system.

The remainder of this paper is organized as follows. Section 2 reviews related works regarding CNN-based and Transformer-based approaches for deepfake detection. Section 3 presents the proposed VERITAS architecture in detail, including CNN blocks, ViT components, and SE blocks. Section 4 discusses the experimental setup, evaluation metrics, comparative results, and ablation study of the model architecture. Finally, Section 5 concludes the paper and suggests future research directions.

2.1 CNN based approach

The architecture in building deepfake detection initially relied heavily on CNN because of its ability to extract spatial features from facial images. One approach that uses CNN is EfficientNetB4 by modifying the attention layer and siamese training [16]. This strategy has been shown to improve performance in detecting facial manipulation by deepfake models. Other studies highlight the challenges in detecting deepfakes with both low and high quality simultaneously. The QAD (Quality-Agnostic Deepfake detection) model was developed with an intra-model collaborative learning approach and maximizes the dependency between feature representations using the Hilbert-Schmidt Independence Criterion (HSIC) [17]. This technique combines Fast Fourier Transform (FFT) and Local Binary Pattern (LBP) to detect traces of manipulation in the texture and frequency domains. Research [18] introduced RealForensics which tries to generalize detection to manipulation that has never been seen before by utilizing original videos of talking faces. Through self-supervised cross-modal learning, this model learns from the natural alignment between visual and audio information to improve generalization capabilities. However, CNN-based models often struggle to generalize across datasets and are prone to overfitting. They focus on superficial visual cues and are easily fooled by high-quality manipulations or compression artifacts.

2.2 Transformer based approach

The increasing need to understand temporal dependencies in deepfake videos has led to Transformer-based architectures being widely applied due to their performance. Research [19] combines Fully Temporal Convolution Network (FTCN) and Temporal Transformer to capture temporal coherence more deeply. The use of FTCN has succeeded in extracting facial motion patterns efficiently using small spatial kernels and large temporal kernels, thus showing its superiority in detecting new types of manipulation. Another study by [20] combined CNN with Vision Transformer (ViT) which was used to detect facial parts such as eyes and nose, and unified predictions from various facial parts using majority voting. This model shows the potential superiority of deepfake detection models built by integrating CNN and Transformer. Furthermore, the GenConViT model [21] comes by combining ConvNeXt and Swin Transformer for visual feature extraction and Autoencoder and Variational Autoencoder to learn latent distributions. This model shows high performance in detecting deepfakes across various datasets, although it still faces challenges in generalizing to data outside the distribution. In addition, there is CViT2 [22] which combines CNN and Vision Transformer in an attention-based detection system. This model shows rapid developments in forensic applications and misinformation tracking. Nevertheless, Transformer-based approaches require large amounts of data for effective training and are computationally intensive. Furthermore, ViT-based models may underperform without sufficient inductive biases when trained on small or imbalanced datasets.

2.3 Challenges and motivations for VERITAS

Although many approaches have been developed, challenges and limitations in terms of efficiency, generalization to new manipulations, and robustness to visual disturbances remain major issues. VERITAS is designed to address these challenges by integrating CNN-Transformer architecture SE-Blocks in an integrated manner. Thus creating a powerful combination of local and global representations for the robustness needs of the detection system. By leveraging the advantages of CNN and Transformer techniques, and applying adaptive regularization and attention strategies, VERITAS is expected to be able to improve efficient, accurate, and robust deepfake detection against disturbances and data distribution variations.

3. Proposed method

In accordance with the results of the analysis related to the need for solutions to the research problems raised in this study, we provide a solution to the problem by developing a deepfake detection model consisting of a combination of CNN with transformers and enhanced with the addition of SE-Block. Fig. 1 shows the design of the deepfake detection architecture that we developed.

VERITAS is developed through three main continuous parts that combine spatial feature extraction from Convolutional Neural Network (CNN), global context modeling from Vision Transformer (ViT), and channel attention module through Squeeze-and-Excitation (SE) Block introduced by [23].

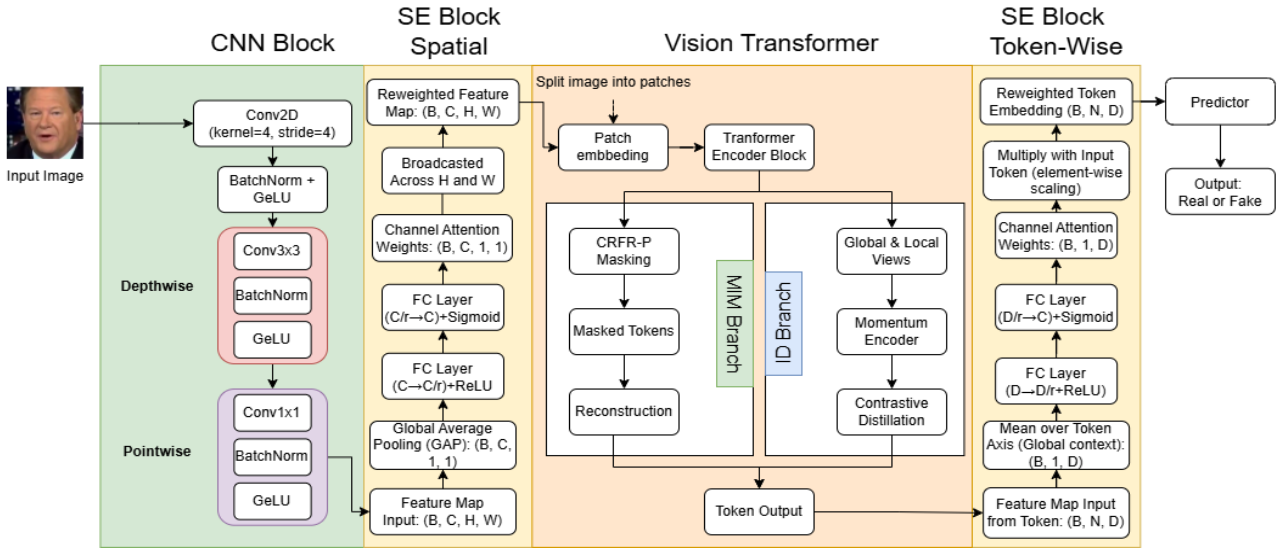


Figure. 1 Proposed Method

In addition, this architectural model is strengthened through two self-learning branches in the transformer part with the aim of improving visual representation in a generalist manner based on the framework of FSFM [24].

3.1 CNN block

The CNN block at the beginning of the VERITAS architecture acts as a feature extractor that focuses on processing local feature representations from images. Where the input image in question contains data in the form of batch values (B), number of channels (RGB), image height (H), and image width (W) so that the input becomes X∈ R^(B×3×H×W)The feature extraction processing circuit in the architecture we designed adopts the ConveXt operation [25] so that it is lighter and more effective in handling tasks with large amounts of data. When X enters this block, the first step is to downsample to reduce the spatial dimension which aims to aggressively reduce computational costs. This downsampling operation involves a convolution layer with a 4x4 kernel and stride 4. Thus producing spatial output sizes H' and W^' based on Eqs. (1) and (2).

$$H' = \left[\frac{H-4}{4}\right] + 1 \quad (1)$$

$$W' = \left[\frac{W-4}{4}\right] + 1 \quad (2)$$

We use an input image size of 224x224, resulting in a spatial output size of H' and W' of 56 and a number of feature channels of 64 defined as $Y \in \mathbb{R}^{B\times 64\times 56\times 56}$. After obtaining these sizes, we continue by normalizing using the batch normalization method and activating using the GELU function. The output of this process then enters the depthwise operation section using Eq. (3).

$$Y(b,c,i,j) = \sum_{m=-k'}^{k'} \sum_{n=-k'}^{k'} K_c(m,n) X(b,c,i+m,j+n) \quad (3)$$

With $k' = k/2$ and k is the kernel size used for the convolution process. $Y(b,c,i,j)$ is the output of the depthwise operation at position (i,j) and channel c for the bth sample. $K_c(m,n)$ is the depthwise filter for channel c that has a size of k×k with (m,n) as the filter index that runs from $-k'$ to $+k'$. $X(b,c,i+m,j+n)$ is the input pixel value at position $(i+m,j+n)$ on channel c for the bth sample. Each output Y produced from this operation is the result of the multiplication and addition of the filter elements K_c with the input X around (i,j). The process on each channel is carried out independently so as to minimize the increase in the number of parameters due to the addition of this CNN block. The operation on this CNN block ends with a pointwise operation that is tasked with combining information between channels at each spatial position using Eq. (4).

$$Y(b,c',i,j) = \sum_{c=1}^{C} W_{c',c} X(b,c,i,j) + b_{c'} \quad (4)$$

Where $Y(b,c',i,j)$ is the output value generated by the pointwise convolution at position (i,j) in channel c' for sample b, $W_{c',c}$ is the weight for the pointwise layer with size 1×1 that connects the input of channel c to channel c', X(b,c,i,j) is the input generated from the depthwise operation in the form of $\mathbb{R}^{B\times C\times H\times W}$, $b_{c'}$ is the bias used for the output of channel c' and c' is the number of output channels from pointwise. The fusion of information from each channel is done by performing a linear combination using the weight $W_{c',c}$, so that an integrated channel space transformation occurs between global and local features.

3.2 SE block spatial

This section consists of three main parts, namely squeeze, excitation, and reweighting with the aim of adaptively recalibrating the channel attention based on the global context of each spatial feature extracted by the CNN block [23]. The output of the CNN block in the form of a 4D spatial tensor denoted as feature F (where $F \in \mathbb{R}^{B\times C\times H\times W}$) is reduced to one vector per channel using the global average pooling (GAP) method [26] in Eq. (5).

$$s_c = \frac{1}{H.W} \sum_{i=1}^{H} \sum_{j=1}^{W} F_{c,i,j}, \quad \forall c \in \{1,\dots,C\} \quad (5)$$

Where s_c represents the average intensity of the c-th channel, H and W are the spatial dimensions of the image, and $F_{c,i,j}$ is the output value of the CNN block. The result of this calculation produces a tensor s (where $s \in \mathbb{R}^{B\times C\times 1\times 1}$). Furthermore, the global squeezed vector (s) enters the excitation section which has two fully-connected layers with each having ReLU and Sigmoid activation functions as shown by Eqs. (6) and (7).

$$z = ReLU\big(Conv1\times 1_{reduce}(s)\big), z \in \mathbb{R}^{B\times C_r\times 1\times 1} \quad (6)$$

$$e = \sigma\left(Conv1 \times 1_{expand}(z)\right), \quad e \in \mathbb{R}^{B\times C\times 1\times 1} \tag{7}$$

With z is the activation result value by the ReLU function, $C_r = C/r$ where r is the reduction ratio, e is the attention vector per channel, and σ is the sigmoid activation function in limiting the output in the range [0,1]. This section serves to calculate the channel attention score nonlinearly in capturing the dependency between channels. After getting the value of e, the next step is to recalibrate the initial feature F by performing channel-wise multiplication on Eq. (8).

$$F_{se} = F \odot e, \qquad F_{se} \in \mathbb{R}^{B\times C\times H\times W} \tag{8}$$

Where $\odot$ is the element-wise broadcast multiplication along $H \times W$ which is done by multiplying each channel c of F by the corresponding scalar value e_c . Thus, SE Block Spatial effectively improves the quality of the local representation of CNN results which will later be projected into tokens through patch embedding operations for the needs of transformer input in the next block.

3.3 Vision transformer block

The ViT block section of the architecture we designed has a function not only to act as a feature encoder, but is also equipped with two self-learning paths that function to maximize the token representation capability of ViT, both locally and globally. The SE Block Spatial (F_{se}) feature results will be converted into token input through the patch embedding process to produce tensor input $X \in \mathbb{R}^{B\times N\times D}$ where B is the number of batches, N is the number of patches, and D is the dimension of the token embedding. After that, the patch embedding input results will enter the transformer encoder, where the tensor input X will go through several layers. The layer starts with Multi-Head Self-Attention (MSA) which allows the model to focus on different parts in one token sequence. Furthermore, the output results from this MSA are forwarded to the feed-forward network (FFN) which is equipped with the GELU activation function. A normalization (LN) layer is added to the MSA and FFN layers so that this process can be formulated as Eqs. (9) and (10).

$$X' = X + MSA(LN(X)) \tag{9}$$

$$X'' = X' + FFN\left(LN(X)\right) \tag{10}$$

Where X is the input of the patch embedding result, X' is the output generated by the MSA layer, and X'' is the output generated by the FFN layer. This encoder process ends by averaging all tokens using mean pooling. The output token from the encoder process is directly used for the self-learning path. The self-learning path in the architecture we developed adopts the FSFM framework [24] so that it consists of two branches of the path, namely the Masked Image Modeling (MIM) Branch and the Identity Distillation (ID) Branch. The MIM Branch focuses on understanding local visual structures through patch reconstruction while the ID Branch learns the global representation of facial identity through contrastive distillation between representations [24]. These two branches maximize the token representation capability of ViT, both locally and globally, to detect deepfakes more accurately and robustly. The results of the self-learning process produce an output token T where $T \in \mathbb{R}^{B\times N\times D}$.

3.4 SE block token-wise

Before entering the predictor section, the output token T from the two self-learning branches (MIM and ID) is first recalibrated so that the model can pay attention to tokens that are more important in the context of deepfake detection. In addition, with T entering the SE Block Token-Wise, it is expected to improve classification performance by strengthening features that truly represent "real" or "fake". Similar to the previous SE Block, this block begins by squeezing using token-level pooling in Eq. (11). The goal is to extract global representations between tokens (not spatial as in CNN).

$$s = \frac{1}{N}\sum_{i=1}^{N} T_{:,i,:} \tag{11}$$

Where s is the average embedding of all tokens in an image representing the global semantic summary ($s \in \mathbb{R}^{B\times 1\times D}$), N is the number of tokens in an image, and T is the input token. Next, s is processed in the excitation section through two fully connected (FC) layers to generate token attention scores per dimension using Eq. (12).

$$a = \sigma(W_2 \cdot ReLU(W_1 \cdot s)) \tag{12}$$

Where a is the attention score value between 0 and 1 ($a \in \mathbb{R}^{B\times 1\times D}$), σ is the sigmoid activation function, W_1 is the down-dimension projection value where $W_1 \in \mathbb{R}^{D\times D_r}$,

Table 1. Notation list

Symbol	Description
B	Batch size (number of input samples)
C	Number of channels in the input image (e.g., 3 for RGB)
H, W	Height and width of the input image
H', W'	Downsampled height and width after convolution (see Eq. 1 and 2)
k	Kernel size for convolution operation
k'	Half kernel size used in depthwise convolution
$X \in \mathbb{R}^{B\times C\times H\times W}$	Input tensor/image
$Y \in \mathbb{R}^{B\times C'\times H'\times W'}$	Output feature map after convolution
$Y(b,c,i,j)$	Value at position (i, j) and channel c for batch b in convolution output
$K_c(m,n)$	Depthwise filter for channel c, with size $k \times k$
$W_{c',c}$	Weight for pointwise convolution mapping channel c to c'
$b_{c'}$	Bias term for output channel c' in pointwise convolution
$F \in \mathbb{R}^{B\times C\times H\times W}$	Output tensor from CNN block (before SE Block)
s_c	Squeezed scalar for channel c via Global Average Pooling (Eq. 5)
$s \in \mathbb{R}^{B\times C\times 1\times 1}$	Squeezed vector over all channels
$z \in \mathbb{R}^{B\times C_r\times 1\times 1}$	Excitation intermediate vector after ReLU (Eq. 6)
C_r	Reduced channel size after first FC layer in SE Block
$e \in \mathbb{R}^{B\times C\times 1\times 1}$	Channel-wise attention weights (Eq. 7)
σ	Sigmoid activation function
δ	ReLU (or GELU) activation function
$F_{se} \in \mathbb{R}^{B\times C\times H\times W}$	Output of spatial SE Block after reweighting (Eq. 8)
N	Number of image patches (tokens) in ViT
D	Dimension of each token embedding
$X \in \mathbb{R}^{B\times N\times D}$	Input to ViT encoder after patch embedding
X'	Output after MSA layer in ViT (Eq. 9)
X''	Output after FFN layer in ViT (Eq. 10)
$T \in \mathbb{R}^{B\times N\times D}$	Output token representation after self-learning branches
$s \in \mathbb{R}^{B\times C\times D}$	Token-level average embedding (Eq. 11)
D^r	Reduced token embedding dimension in SE Token-Wise block
$W_1 \in \mathbb{R}^{D\times D_r}$	Weight matrix of first FC layer in SE Token-Wise block
$W_2 \in \mathbb{R}^{D_r\times D}$	Weight matrix of second FC layer (expansion back to D)
$a \in \mathbb{R}^{B\times 1\times D}$	Token-wise attention vector after excitation (Eq. 12)
$Z_{se} \in \mathbb{R}^{B\times N\times D}$	Recalibrated tokens using token-wise attention (Eq. 13)
$\hat{y}$	Final output score (real or fake) from prediction layer
$\odot$	Element-wise or broadcast multiplication operator

W_2 is the back projection to the original dimension (excitation) where $W_2 \in \mathbb{R}^{D_r\times D}$, D_r is the reduced token embedding dimension where $D_r = \frac{D}{r}$, and r is the reduction ratio. Through these two FC layers, the token-wise Block SE is able to learn the important part of each dimension of the token embedding and provide focused reinforcement. After obtaining the attention value from the excitation result, all T tokens are recalibrated through Eq. (13).

$$Z_{se} = T \odot a \tag{13}$$

Where Z_{se} is the token recalibrated by a and $\odot$ is the broadcast multiplication between tokens and attention vectors. The output of Z_{se} is then sent to the predictor section which acts as the final classification component that changes the representation of the token processed by Vision Transformer into a probability score whether the input is a real face or a digitally manipulated face (deepfake).

The use of SE blocks in both the spatial (CNN) and token (ViT) domains is intended to perform hierarchical attention calibration. Spatial SE in the CNN block enhances local features relevant to forgery artifacts (e.g., texture or edge inconsistencies), while token-wise SE in ViT emphasizes global semantic cues (e.g., identity mismatch). This dual strategy acts as a progressive filter first refining low-level noise, then reinforcing high-level semantics thus reducing overfitting and enhancing generalization. By applying channel-wise weighting (in Eq. (8)) and token-wise recalibration (in Eq. (13)), the model avoids feature redundancy and learns complementary information from both domains. This nested attention mechanism ensures more robust and discriminative representations for deepfake detection.

All notations are consistent throughout the equations to ensure clarity in representing the intermediate calculations of the VERITAS architecture as shown in Table 1.

4. Experiments

Given that VERITAS combines CNN, vision transformer, and SE block components into one unit, this causes uncertainty in selecting the right optimizer method. Therefore, we conducted a small-scale test by training the VERITAS model using several optimizer methods. This optimizer selection test uses the FaceForensics++ dataset [27] with an image size of 224×224, a batch size of 64, an epoch number of 10, a learning rate of 25e-5, and without using a scheduler. The test results are shown in Table 2.

The experimental results in Table 2 show that the implementation of AdamW optimizer on the VERITAS architecture provides the best performance compared to other methods. This is in line with various previous studies that prove the suitability of AdamW for Transformer-based models due to the use of weight decay that is separate from parameter updates [28], so that it can improve model regularization and generalization [29]–[31]. To clarify the results of the effectiveness of each optimizer, Fig. 2 shows a decrease in the loss value and an increase in the AUC validation value.

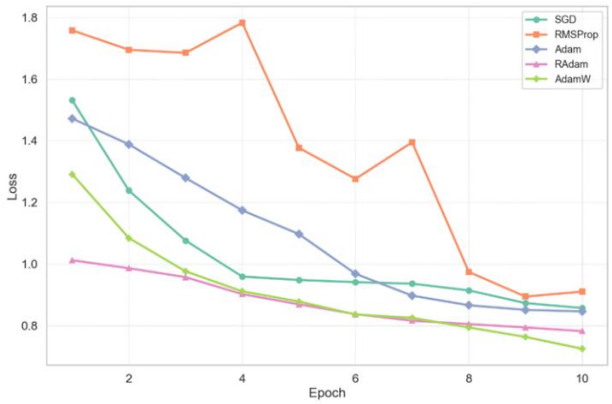
However, the convergence speed of AdamW is still below RAdam. This is evidenced by RAdam reaching convergence in the 2nd epoch. In addition, in Fig. 2(a), RAdam has a more stable loss reduction compared to the reduction in AdamW. Although RAdam has quite good speed and stability, to train the VERITAS model we use the AdamW optimizer because the final results are more optimal.

The model is trained using the ViT-B/16 backbone, using the VF2 ViT-B model pretrain [24], and input images of size 224×224. During training, 75% of patches are masked in the Masked Image Modeling (MIM) branch to force the model to reconstruct missing tokens, while the Identity Distillation (ID) branch uses contrastive learning with augmented views of the same image, leveraging a temperature of 0.1. To enhance reproducibility, we provide the following simplified pseudo-code based on FSFM framework [24] in Table 3.

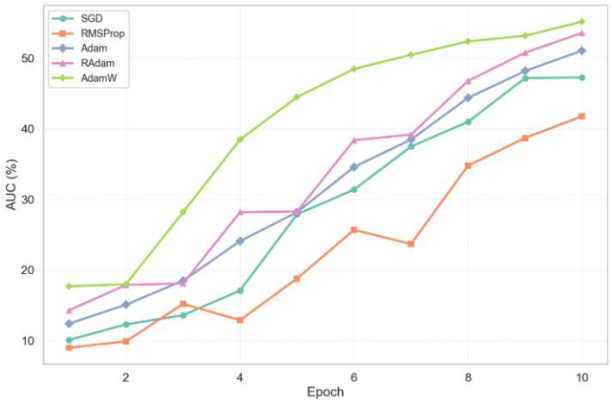
Training is performed with a batch size of 64, learning rate of 2.5e-5, and 100 epochs with 5 warm-up epochs, using FaceForensics++ (FF++, c23/HQ version) [27] as training data. This dataset comes from video data that is converted into images per frame with a total of 127848. We use a ratio distribution composition of 70% for training data, 15% for test data, and 15% for validation data with random shuffling. The model is evaluated on the unseen datasets CelebDF-v2 (CDFV2) [32] and Wild Deep-fake (WDF) [33]. using the Area Under Curve (AUC) metric at both the frame-level and video-level.

Table 2. Effectiveness of each optimizer on the VERITAS architecture

Optimizer	AUC	Final Loss	Convergence Speed
SGD	47.3%	0.857	Slow (takes more than 10+ epochs)
RMSProp	41.8%	0.910	Up and Down
Adam	53.1%	0.846	Stable at the 4th epoch
RAdam	53.6%	0.782	Stable at the 2nd epoch
AdamW	55.2%	0.725	Stable at the 3rd epoch



(a)



(b)

Figure. 2: (a) Reduction in loss value of each optimizer method and (b) Increase in AUC value (%) of each optimizer method

The evaluation results are shown in Tables 4 and Table 5.

In evaluating the performance of VERITAS, we conducted cross-dataset testing on Celeb-DF v2 and WildDeepfake at the frame and video levels. The results show that VERITAS consistently achieves the highest AUC compared to previous methods with 87.45% and 88.30% at the frame level and 95.56% and 87.75% at the video level.

Table 3. Pseudo-Code for Self-Supervised Learning Branches

```
Algorithm 1: Self-Supervised Branches in VERITAS

Input: Image batch X
1. Patch Embed: Z ← PatchEmbed(X) # Z ∈ R^{B×N×D}
2. MIM Branch:
 a. Mask 75% of tokens:
    Z_masked ← mask(Z, ratio=0.75)
 b. Reconstruct tokens:
    Z_recon ← Decoder(Z_masked)
 c. Compute loss:
    L_MIM ← MSE(Z_recon, Z_original)

3. ID Branch:
  a. Generate two augmentations:
     Z1, Z2 ← augment(Z)
  b. Normalize tokens:
     Z1_norm, Z2_norm ← normalize(Z1), normalize(Z2)
  c. Compute contrastive loss:
     L_ID ← NT_Xent(Z1_norm, Z2_norm, τ=0.1)

4. Classifier:
   y_hat ← Predictor(Z)
5. Supervised loss:
   L_cls ← Cross_Etropy(y_hat, y_true)

Output: Total loss L_total ← L_cls + L_MIM + L_ID
```

This improvement is due to the VERITAS architectural design that combines local feature extraction through CNN blocks and global relation modeling through ViT which is strengthened by the attention calibration mechanism using SE Block in the spatial and token domains. The dual-branch strategy with Masked Image Modeling and Identity Distillation in the transformer block section that adopts the FSFM framework [24] further enhances the semantic representation of the model. Thus, VERITAS is able to detect deepfakes more effectively in data distributions that are different from the training data.

We provide an ablation study by comparing several variants of the VERITAS model component fractions based on the Area Under Curve (AUC) performance and confusion metric on the FaceForensics++ (FF++) validation data [27]. The compared models include the baseline FSFM [24], Transformer combined with SE Block in the token domain (TF+SE), CNN followed by Transformer without SE Block (CNN+TF), CNN with SE Block in the spatial domain before Transformer (CNN+SE+TF), CNN with Transformer combined

Table 4. Cross-dataset evaluation on deepfake detection for frame-level comparison

Model	Train set	AUC	
		CDFV2	WDF
OST [34]	FF++	74.80%	-
FInfer [35]	FF++	70.60%	69.46%
PEL [36]	FF++	69.18%	67.39%
SLADD [37]	SD	79.70%	-
RECCE [38]	FF++	68.71%	64.31%
UIA-ViT [39]	FF++	82.41%	-
UAL [40]	FF++	82.84%	70.13%
NoiseDF [41]	FF++	75.89%	-
GS [42]	FF++	84.97%	-
UCF [43]	FF++	82.40%	-
SFDG [44]	FF++	75.83%	69.27%
IID [45]	FF++	83.80%	-
LSDA [46]	FF++	83.00%	-
FSFM [24]	FF++	85.05%	85.26%
VERITAS	**FF++**	**87.45%**	**88.30%**

Table 5. Cross-dataset evaluation on deepfake detection for video-level comparison

Model	Train Set	AUC	
		CDFV2	WDF
SBIs [47]	SD	93.18%	-
RealForensics [18]	FF++	86.90%	-
HCIL [48]	FF++	79.00%	-
SeeAble [49]	SD	87.30%	
Exploring Temporal Coherence [19]	FF++	86.9%	-
TALL [50]	SD	90.79%	-
AUNet [51]	SD	92.77%	-
SLF [52]	FF++	89.00%	-
MLR [53]	FF++	91.56%	73.41%
LAA-Net/BI [54]	SD	86.28%	57.13%
LAA-Net/SBI [54]	SD	95.40%	80.03%
LSDA [46]	FF++	91.10%	-
FPG [55]	SD	94.49%	-
NACO [56]	FF++	89.50%	-
FSFM [24]	FF++	91.44%	86.96%
VERITAS	**FF++**	**95.56%**	**87.75%**

Table 6. The results of the ablation study on the FF++ validation data were measured using AUC (%) at the frame-level and video-level and confusion metric.

Model Variant	AUC Frame-Level (%)	AUC Video-Level (%)	Precision (%)	Recall (%)	F1-Score (%)
FSFM (Baseline) [24]	76.39	82.31	80.15	78.63	79.38
TF+SE	72.43	79.17	68.51	74.39	71.33
CNN+TF	74.31	80.58	70.40	77.83	73.93
CNN+SE+TF	76.83	83.29	72.17	80.48	76.10
CNN+TF+SE	77.92	85.24	74.92	82.33	78.45
CNN+SE+TF +SE	**83.79**	**87.15**	**80.02**	**88.61**	**84.10**

with SE Block in the token domain (CNN+TF+SE), and the complete configuration of CNN+SE+TF+SE (full VERITAS architecture). The results of the study are shown in Table 6.

The results in Table 6 show that the baseline FSFM [24] obtained AUC values of 76.39% for frame-level and 82.31% for video-level. The precision of this baseline model is 80.15%, and recall is 78.63%, resulting in an F1-score of 79.38%. This model performs decently but still leaves room for improvement, especially in the recall, reflecting its ability to identify manipulated instances correctly. When using only Transformer with SE Block (TF+SE), the performance of the model decreased to an AUC of 72.43% at the frame-level and 79.17% at the video-level. The precision dropped to 68.51%, and recall slightly increased to 74.39%, resulting in an F1-score of 71.33%. This indicates that the absence of local features from CNN weakens the model's ability to capture manipulation artifacts effectively, leading to a reduction in both precision and recall. Adding CNN features to the Transformer (CNN+TF) improves the model's performance to AUC values of 74.31% and 80.58% at the frame- and video-levels, respectively. The precision increases to 70.40%, and recall improves to 77.83%, leading to a higher F1-score of 73.93%. The inclusion of CNN features helps the model better detect local manipulation artifacts, although there's still a gap in the model's recall performance compared to other configurations. By inserting the SE Block before the Transformer (CNN+SE+TF), the model achieves AUC values of 76.83% (frame-level) and 83.29% (video-level). The precision increases to 72.17%, and recall improves significantly to 80.48%, with an F1-score of 76.10%.

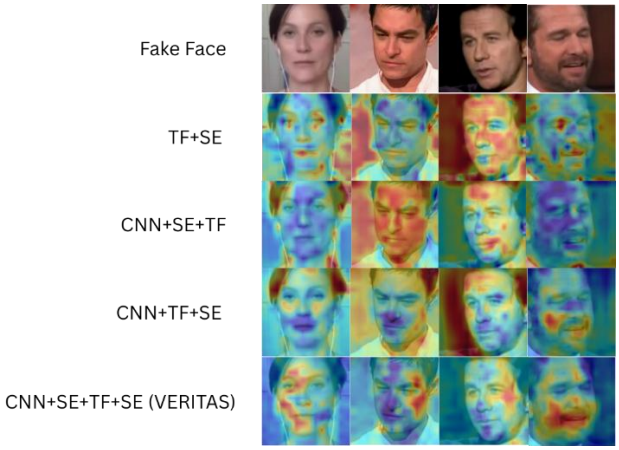


Figure. 3 Visualization of deepfake prediction results using various VERITAS variants on the CDFV2 dataset

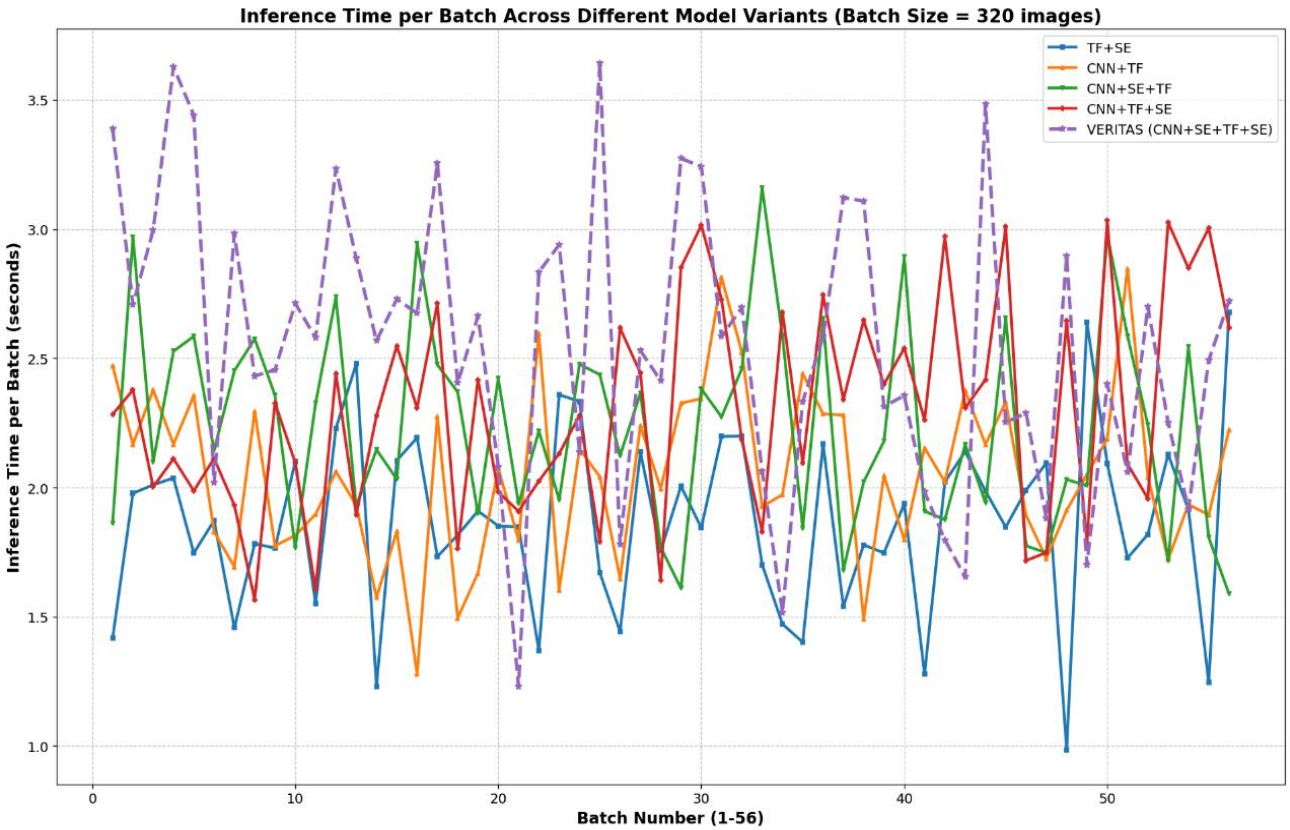


Figure. 4 Comparison of inference time of each VERITAS model variant

This demonstrates that spatial feature calibration with SE Block enhances both the model's ability to focus on important local features and its overall performance. Placing the SE Block after the Transformer (CNN+TF+SE) further improves performance with an AUC of 77.92% for frame-level and 85.24% for video-level. Precision increases to 74.92%, and recall increases to 82.33%, resulting in an F1-score of 78.45%. This configuration benefits from token-wise attention that enriches semantic information, improving both precision and recall. The full configuration (CNN+SE+TF+SE) produces the best performance, with AUC values of 83.79% and 87.15%. The precision reaches 80.02%, and recall improves significantly to 88.61%, leading to the highest F1-score of 84.10%. This final configuration demonstrates the importance of applying multilevel attention to both spatial and token domains.

The model achieves the most robust representation for deepfake detection, reflecting its ability to generalize across different types of manipulations and conditions. In addition, in terms of qualitative quality, the detection produced by the full configuration shows better accuracy than other variants. This reinforces the importance of combining both spatial and token attention mechanisms for improved deepfake detection. This is evidenced in Fig. 3 which shows a comparison of deepfake detection results visualized using the GradCam method [57]. However, due to the addition of new blocks, this causes the VERITAS inference time to be slightly slower as shown in Fig. 4.

5. Conclusion

This study proposes and evaluates VERITAS, a deepfake detection architecture that combines CNN, Vision Transformer (ViT), and adaptive attention mechanisms through Squeeze-and-Excitation (SE) Blocks in both spatial and token domains. Experimental results demonstrate that VERITAS significantly outperforms various state-of-the-art methods, achieving AUC values of 83.79% (frame-level) and 87.15% (video-level), compared to the FSFM baseline's AUC of 76.39% and 82.31%, respectively. In addition, precision (80.02%), recall (88.61%), and F1-score (84.10%) results further validate the superior detection performance of VERITAS, especially in identifying manipulated faces while maintaining a low false-positive rate. The ablation study confirms that the adaptive application of SE Block at both spatial and token levels significantly enhances the model's ability to detect facial manipulations by improving semantic representation and feature refinement.

The integration of two self-supervised learning branches Masked Image Modelling (MIM) and Identity Distillation (ID) has shown to improve the model's generalization and resilience across different data distributions, leading to more accurate and robust detection of deepfake content. Despite these advancements, VERITAS still faces challenges in inference time efficiency due to the complexity of its architecture. Future work should focus on optimizing the model for real-time deployment, perhaps through lightweight transformer techniques, knowledge distillation, or more efficient attention mechanisms such as sparse attention. Furthermore, expanding the testing to include multimodal audio-visual manipulation and developing a continuous learning-based detection system would strengthen the adaptability of VERITAS and make it more suitable for deployment in dynamic real-world environments. In addition, model testing through perturbation testing such as compression artifacts, frame drops, or lighting variations is needed to test the model's robustness to various real-world conditions.

Conflicts of interest

The Author declared that they have no competing interests.

Author contributions

Alam Rahmatulloh compiled, designed, and collected research data, analyzed data, experiments and testing and contributed to writing the manuscript. Herman Dwi Surjono and Fatchul Arifin were the research supervisors and mentors. Irfan Darmawan and Nia Ambarsari contributed to validation of test data and data analysis. All authors have approved the final version.

Acknowledgments

This research is funded by the Directorate of Research and Community Service, Directorate General of Research and Development (DRTPM), Ministry of Higher Education, Science, and Technology (Kemdiktisaintek) of the Republic of Indonesia in 2025.

References

[1] A. Rahmatulloh, H. D. Surjono, F. Arifin, and G. F. Nugraha, "Exploring Deepfake Models for Image Translation: A Systematic Review of Current Techniques and Future Directions", *Social Science Research Network (SSRN)*, 2024, doi: 10.2139/ssrn.5042547.

[2] S. A. Khan, and D.-T. Dang-Nguyen, "Deepfake Detection: Analyzing Model Generalization Across Architectures, Datasets, and Pre-Training Paradigms", *IEEE Access*, Vol. 12, pp. 1880-1908, 2024, doi: 10.1109/ACCESS.2023.3348450.

[3] D. A. Coccomini, R. Caldelli, F. Falchi, and C. Gennaro, "On the Generalization of Deep Learning Models in Video Deepfake Detection", *Journal of Imaging*, Vol. 9, No. 5, p. 89, 2023, doi: 10.3390/jimaging9050089.

[4] L. Y. Gong, X. J. Li, and P. H. J. Chong, "Swin-Fake: A Consistency Learning Transformer-Based Deepfake Video Detector", *Electronics*, Vol. 13, No. 15, p. 3045, 2024, doi: 10.3390/electronics13153045.

[5] B. Ghita, I. Kuzminykh, A. Usama, T. Bakhshi, and J. Marchang, "Deepfake Image Detection Using Vision Transformer Models", In: *Proc. of 2024 IEEE International Black Sea Conference on Communications and Networking, BlackSeaCom 2024*, pp. 332-335, 2024, doi: 10.1109/BLACKSEACOM61746.2024.10646310.

[6] S. A. Khan, and D.-T. Dang-Nguyen, "Hybrid Transformer Network for Deepfake Detection", In: *Proc. of International Conference on Content-based Multimedia Indexing*, pp. 8-14, 2022, doi: 10.1145/3549555.3549588.

[7] Z. Tan, Z. Yang, C. Miao, and G. Guo, "Transformer-Based Feature Compensation and Aggregation for DeepFake Detection", *IEEE Signal Processing Letters*, Vol. 29, pp. 2183-2187, 2022, doi: 10.1109/LSP.2022.3214768.

[8] P. M. Thuan, B. T. Lam, and P. D. Trung, "DSViT: An Enhanced Transformer Model for Deepfake Detection", *Journal of Science and Technology on Information Security*, No. 2, pp. 17-28, 2024, doi: 10.54654/ISJ.V2I22.1055.

[9] T. Wang, H. Cheng, K. P. Chow, and L. Nie, "Deep Convolutional Pooling Transformer for Deepfake Detection", *ACM Transactions on Multimedia Computing, Communications and Applications*, Vol. 19, No. 6, 2022, doi: 10.1145/3588574.

[10] D. Huang, and Y. Zhang, "Learning Meta Model for Strong Generalization Deepfake Detection", In: *Proc. of 2024 International Joint Conference on Neural Networks (IJCNN)*, pp. 1-8, 2024, doi: 10.1109/IJCNN60899.2024.10651482.

[11] H. She, Y. Hu, B. Liu, J. Li, and C. T. Li, "Using Graph Neural Networks to Improve Generalization Capability of the Models for Deepfake Detection", *IEEE Transactions on Information Forensics and Security*, 2024, doi: 10.1109/TIFS.2024.3451356.

[12] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation Networks", In: *Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 7132-7141, 2018, doi: 10.1109/CVPR.2018.00745.

[13] Z. Akhtar, "Deepfakes Generation and Detection: A Short Survey", *Journal of Imaging*, Vol. 9, No. 1, 2023, doi: 10.3390/jimaging9010018.

[14] A. Heidari, N. J. Navimipour, H. Dag, and M. Unal, "Deepfake detection using deep learning methods: A systematic and comprehensive review", *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Vol. 14, No. 2, pp. 1-45, 2024, doi: 10.1002/widm.1520.

[15] L. Y. Gong, and X. J. Li, "A Contemporary Survey on Deepfake Detection: Datasets, Algorithms, and Challenges", *Electronics (Switzerland)*, Vol. 13, No. 3, 2024, doi: 10.3390/electronics13030585

[16] N. Bonettini, E. D. Cannas, S. Mandelli, L. Bondi, P. Bestagini, and S. Tubaro, "Video Face Manipulation Detection Through Ensemble of CNNs", In: *Proc. of 2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 5012-5019, 2021, doi: 10.1109/ICPR48806.2021.9412711.

[17] B. M. Le, and S. S. Woo, "Quality-Agnostic Deepfake Detection with Intra-model Collaborative Learning", In: *Proc. of 2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22321-22332, 2023, doi: 10.1109/ICCV51070.2023.02045.

[18] A. Haliassos, R. Mira, S. Petridis, and M. Pantic, "Leveraging Real Talking Faces via Self-Supervision for Robust Forgery Detection", In: *Proc. of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14930-14942, 2022, doi: 10.1109/CVPR52688.2022.01453.

[19] Y. Zheng, J. Bao, D. Chen, M. Zeng, and F. Wen, "Exploring Temporal Coherence for More General Video Face Forgery Detection", In: *Proc. of 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15024-15034, 2021, doi: 10.1109/ICCV48922.2021.01477.

[20] A. H. Soudy, *et al.*, "Deepfake detection using convolutional vision transformers and convolutional neural networks", *Neural Computing and Applications*, 2024, doi: 10.1007/s00521-024-10181-7.

[21] D. W. Deressa, H. Mareen, P. Lambert, S. Atnafu, Z. Akhtar, and G. Van Wallendael, "GenConViT: Deepfake Video Detection Using Generative Convolutional Vision Transformer", *arXiv Preprint, arXiv:2307.07036*, 2023.

[22] D. W. Deressa, P. Lambert, G. Van Wallendael, S. Atnafu, and H. Mareen, "Improved Deepfake Video Detection Using Convolutional Vision Transformer", In: *Proc. of 2024 IEEE Gaming, Entertainment, and Media Conference, GEM 2024*, 2024, doi: 10.1109/GEM61861.2024.10585593.

[23] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation Networks", In: *Proc. of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7132-7141, 2018, doi: 10.1109/CVPR.2018.00745.

[24] G. Wang, F. Lin, T. Wu, Z. Liu, Z. Ba, and K. Ren, "FSFM: A Generalizable Face Security Foundation Model via Self-Supervised Facial Representation Learning", *arXiv Preprint, arXiv:2412.12032*, 2024.

[25] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s", In: *Proc. of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11966-11976, 2022, doi: 10.1109/CVPR52688.2022.01167.

[26] M. Lin, Q. Chen, and S. Yan, "Network In Network", *arXiv Preprint, arXiv:1312.4400*, 2013.

[27] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, "FaceForensics++: Learning to Detect Manipulated Facial Images", In: *Proc. of 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1-11, 2019, doi: 10.1109/ICCV.2019.00009.

[28] I. Loshchilov, and F. Hutter, "Decoupled Weight Decay Regularization", *arXiv Preprint, arXiv:1711.05101*, 2017.

[29] A. Dosovitskiy, *et al.*, "An Image is Worth 16X16 Words: Transformers for Image Recognition at Scale", *arXiv Preprint, arXiv 22010.11929*, 2021, doi: 10.48550/arXiv.2010.11929.

[30] A. El-Nouby, N. Neverova, I. Laptev, and H. Jégou, "Training Vision Transformers for Image Retrieval", *arXiv Preprint, arXiv:2102.05644*, 2021.

[31] X. Chen, C.-J. Hsieh, and B. Gong, "When Vision Transformers Outperform ResNets without Pre-training or Strong Data Augmentations", *arXiv Preprint, arXiv:2106.01548*, 2021.

[32] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics", In: *Proc. of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3204-3213, 2020, doi: 10.1109/CVPR42600.2020.00327.

[33] B. Zi, M. Chang, J. Chen, X. Ma, and Y. G. Jiang, "WildDeepfake: A Challenging Real-World Dataset for Deepfake Detection", In: *Proc. of the 28th ACM International Conference on Multimedia*, pp. 2382-2390, 2020, doi: 10.1145/3394171.3413769.

[34] L. Chen, Y. Zhang, Y. Song, J. Wang, and L. Liu, "OST: Improving Generalization of DeepFake Detection via One-Shot Test-Time Training", *Advances in Neural Information Processing Systems*, Vol. 35, pp. 24597-24610, 2022.

[35] J. Hu, X. Liao, J. Liang, W. Zhou, and Z. Qin, "FInfer: Frame Inference-Based Deepfake Detection for High-Visual-Quality Videos", In: *Proc. of the AAAI Conference on Artificial Intelligence*, Vol. 36, No. 1, pp. 951-959, 2022, doi: doi.org/10.1609/aaai.v36i1.19978.

[36] Q. Gu, S. Chen, T. Yao, Y. Chen, S. Ding, and R. Yi, "Exploiting Fine-grained Face Forgery Clues via Progressive Enhancement Learning", In: *Proc. of the AAAI Conference on Artificial Intelligence*, pp. 735-743, 2022.

[37] L. Chen, Y. Zhang, Y. Song, L. Liu, and J. Wang, "Self-supervised Learning of Adversarial Example: Towards Good Generalizations for Deepfake Detection", In: *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18689-18698, 2022, doi: 10.1109/CVPR52688.2022.01815.

[38] J. Cao, C. Ma, T. Yao, S. Chen, S. Ding, and X. Yang, "End-to-End Reconstruction-Classification Learning for Face Forgery Detection", In: *Proc. of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4103-4112, 2022, doi: 10.1109/CVPR52688.2022.00408.

[39] W. Zhuang, *et al.*, "UIA-ViT: Unsupervised Inconsistency-Aware Method based on Vision Transformer for Face Forgery Detection", In: *Proc. of European Conference on Computer Vision*, pp. 391-407, 2022.

[40] Y. Wu, X. Song, J. Chen, and Y.-G. Jiang, "Generalizing Face Forgery Detection via Uncertainty Learning", In: *Proc. of the 31st ACM International Conference on Multimedia*, pp. 1759-1767, 2023, doi: 10.1145/3581783.3612102.

[41] T. Wang, and K. P. Chow, "Noise Based Deepfake Detection via Multi-Head Relative-Interaction", In: *Proc. of the AAAI Conference on Artificial Intelligence*, Vol. 37, No. 12, pp. 14548-14556, 2023, doi: doi.org/10.1609/aaai.v37i12.26701.

[42] Y. Guo, C. Zhen, and P. Yan, "Controllable Guide-Space for Generalizable Face Forgery Detection", In: *Proc. of 2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20761-20770, 2023, doi: 10.1109/ICCV51070.2023.01903.

[43] Z. Yan, Y. Zhang, Y. Fan, and B. Wu, "UCF: Uncovering Common Features for Generalizable Deepfake Detection", In: *Proc. of 2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22355-22366, 2023, doi: 10.1109/ICCV51070.2023.02048.

[44] Y. Wang, K. Yu, C. Chen, X. Hu, and S. Peng, "Dynamic Graph Learning with Content-guided Spatial-Frequency Relation Reasoning for Deepfake Detection", In: *Proc. of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7278-7287, 2023, doi: 10.1109/CVPR52729.2023.00703.

[45] B. Huang *et al.*, "Implicit Identity Driven Deepfake Face Swapping Detection", In: *Proc. of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4490-4499, 2023, doi: 10.1109/CVPR52729.2023.00436.

[46] Z. Yan, Y. Luo, S. Lyu, Q. Liu, and B. Wu, "Transcending Forgery Specificity with Latent Space Augmentation for Generalizable Deepfake Detection", In: *Proc. of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8984-8994, 2024, doi: 10.1109/CVPR52733.2024.00858.

[47] K. Shiohara, and T. Yamasaki, " Detecting Deepfakes with Self-Blended Images ", In: *Proc. of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* , pp. 18699-18708, 2022, doi: 10.1109/CVPR52688.2022.01816.

[48] T. and C. Y. and D. S. and M. L. Gu Zhihao and Yao, "Hierarchical Contrastive Inconsistency Learning for Deepfake Video Detection", In: *Proc. of Computer Vision - ECCV 2022*, pp. 596-613, 2022.

[49] N. Larue, N.-S. Vu, V. Struc, P. Peer, and V. Christophides, "SeeABLE: Soft Discrepancies and Bounded Contrastive Learning for Exposing Deepfakes", In: *Proc. of 2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20954-20964, 2023, doi: 10.1109/ICCV51070.2023.01921.

[50] Y. Xu, J. Liang, G. Jia, Z. Yang, Y. Zhang, and R. He, "TALL: Thumbnail Layout for Deepfake Video Detection ", In: *Proc. of 2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22601-22611, 2023, doi: 10.1109/ICCV51070.2023.02071.

[51] W. Bai, Y. Liu, Z. Zhang, B. Li, and W. Hu, "AUNet: Learning Relations Between Action Units for Face Forgery Detection", In: *Proc. of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24709-24719, 2023, doi: 10.1109/CVPR52729.2023.02367.

[52] J. Choi, T. Kim, Y. Jeong, S. Baek, and J. Choi, "Exploiting Style Latent Flows for Generalizing Deepfake Video Detection", In: *Proc. of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1133-1143, 2024, doi: 10.1109/CVPR52733.2024.00114.

[53] C.-Y. Hong, Y.-C. Hsu, and T.-L. Liu, "Contrastive Learning for DeepFake Classification and Localization via Multi-Label Ranking", In: *Proc. of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17627-17637, 2024, doi: 10.1109/CVPR52733.2024.01669.

[54] D. NGUYEN *et al.*, "LAA-Net: Localized Artifact Attention Network for Quality-Agnostic and Generalizable Deepfake Detection", In: *Proc. of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17395-17405, 2024, doi: 10.1109/CVPR52733.2024.01647.

[55] R. Xia *et al.*, "Advancing Generalized Deepfake Detector with Forgery Perception Guidance", In: *Proc. of the 32nd ACM International Conference on Multimedia*, pp. 6676-6685, 2024, doi: 10.1145/3664647.3680713.

[56] D. Zhang, Z. Xiao, S. Li, F. Lin, J. Li, S. Ge, "Learning Natural Consistency Representation for Face Forgery Video Detection", *Computer Vision - ECCV 2024*, pp. 407-424, 2025.

[57] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization", In: *Proc. of 2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 618-626, 2017, doi: 10.1109/ICCV.2017.74.