ABSTRACT

The proliferation of deepfake technologies has raised significant concerns in digital media integrity, necessitating robust detection systems capable of handling diverse and sophisticated manipulations. Existing approaches range from task-specific models that excel in identifying particular artifacts to generalizable methods leveraging pre-trained foundation models. However, a gap remains in achieving both high accuracy and adaptability across varied datasets and generation techniques.

This study presents a novel framework for deepfake detection that integrates a frozen CLIP vision encoder with multi-scale feature adapters and learnable class-specific context vectors. By refining patch-level features using multi-scale convolutions and augmenting semantic representations with context-aware enhancements, the proposed method captures both local artifacts and global semantic inconsistencies. The system combines the strengths of handcrafted and foundation model-based approaches, ensuring task-specific optimization without sacrificing scalability.

The framework is evaluated on benchmark datasets, including FaceForensics++, DeepFakeDetection, and CELEB-DF, covering a broad spectrum of deepfake generation techniques and quality levels. Results demonstrate that the proposed approach achieves superior performance and generalization compared to existing methods, offering a robust, scalable, and adaptable solution to the challenges of deepfake detection. This work underscores the potential of hybrid strategies that blend the precision of task-specific insights with the versatility of large-scale pre-trained models.

INTRODUCTION

The rapid advancements in generative adversarial networks (GANs) and deep learning techniques have made it increasingly challenging to distinguish between real and manipulated media, particularly in the context of deepfakes. Deepfakes, which involve the creation of highly realistic synthetic images or videos, pose significant ethical, social, and security concerns, necessitating the development of robust detection systems. Traditional detection methods often rely on task-specific architectures, which can effectively target certain types of manipulations but lack the generalization capability required to handle the diversity and complexity of modern deepfake generation techniques.

Recent efforts in deepfake detection have introduced both handcrafted feature extraction models and generalizable approaches leveraging large-scale pre-trained foundation models. Handcrafted models like Mesonet focus on specific visual and frequency-domain features to identify artifacts indicative of manipulation. While effective in controlled scenarios, these models often struggle with cross-dataset generalization and scalability. On the other hand, foundation models, such as CLIP and its variants, exploit pre-trained embeddings that are versatile across multiple domains, offering significant potential for scalable and adaptable detection systems.

In this study, we propose a novel methodology that combines the strengths of foundation models with task-specific adaptations to address the challenges of deepfake detection. Our approach integrates a frozen CLIP vision encoder with multi-scale feature adapters and learnable class-specific context vectors, enabling effective extraction of both local and global features. By enhancing patch-level representations with multi-scale convolutions and incorporating semantic information through learnable context vectors, our framework achieves a fine balance between domain-specific optimization and generalization across diverse datasets.

To evaluate the effectiveness and robustness of the proposed method, we utilize a comprehensive set of benchmarks, including FaceForensics++, DeepFakeDetection, and CELEB-DF, which encompass a wide range of deepfake generation techniques and quality levels. This work aims to advance the state of the art in deepfake detection by leveraging hybrid strategies that integrate the precision of handcrafted feature extraction with the scalability of foundation models.

LITERATURE SURVEY

Various techniques have been developed to detect deepfakes, ranging from task-specific standalone models to more generalizable approaches utilizing large-scale foundation models. Mesonet is a model designed to analyze mesoscopic-level features, which exist between macroscopic (semantic) and microscopic (artifact-level) representations. This intermediate analysis helps address limitations in both macro and micro feature extraction. The architecture of Mesonet is lightweight, employing a small number of convolutional layers to efficiently extract relevant features. Performance improves when standard convolutional layers are replaced with Inception modules, as these capture multi-scale spatial information and highlight subtle patterns indicative of manipulation. While initially developed for video-based detection, Mesonet has demonstrated its utility in image-based deepfake tasks, offering a balance of accuracy and computational efficiency.

Another approach is the DCT classifier, which focuses on frequency-domain features to differentiate real and fake images. This method is inspired by the presence of visible artifacts, such as grid-like patterns, in the frequency spectrum of GAN-generated images. The discrete cosine transform (DCT) is applied to images to extract frequency-domain features, which are then log-scaled for better representation. A logistic regression classifier is trained on these features to identify manipulations. The frequency-domain representation is inherently robust to spatial transformations, contributing to better generalization. While computationally efficient and effective against GAN-based artifacts, this method may struggle to detect manipulations in high-resolution or advanced deepfake techniques that minimize frequency-domain anomalies.

Foundation models like UnivCLIP offer a different perspective by leveraging pre-trained large-scale architectures for deepfake detection. UnivCLIP uses the CLIP:ViT-L/14 model, trained on 400 million image-text pairs, to extract embeddings from its frozen layers. These embeddings are then used in a linear classifier fine-tuned for deepfake detection. UnivCLIP highlights the surprising effectiveness of general-purpose embeddings, which are not explicitly trained for deepfake tasks, in achieving high generalization performance. The simplicity of this approach, combined with its scalability across diverse datasets, represents a significant shift in methodology from handcrafted feature extraction to generalizable embedding-based learning. However, this reliance on pre-trained features can limit task-specific optimizations.

Prompt-tuned CLIP represents another evolution in transfer learning by optimizing learnable prompts for task-specific objectives. Context Optimization (CoOp), a prompt-tuning strategy, appends learnable vectors to context words in textual prompts while keeping the CLIP encoders frozen. These learnable vectors are trained to align with downstream tasks like deepfake detection, with the optimization guided by cross-entropy loss. Experiments with the number of context tokens show that appending them at the front of class labels yields superior results. Prompt tuning is computationally efficient, requiring minimal fine-tuning, and adapts seamlessly to downstream tasks. However, its performance is sensitive to hyperparameters such as the number of context tokens and initialization strategies for the prompts.

Standalone models like Mesonet and DCT classifiers emphasize task-specific feature extraction, excelling in targeted detection scenarios but facing challenges in generalization. In contrast, foundation models like UnivCLIP and Prompt-tuned CLIP prioritize scalability and generalization by leveraging pre-trained embeddings, enabling rapid adaptation to new datasets and tasks. These advancements collectively represent a shift toward hybrid strategies that integrate task-specific insights with the versatility of large-scale pre-trained models.

DATASETS

To evaluate the generalizability of our method across state-of-the-art deepfake generation techniques, we selected datasets that comprehensively cover the major deepfake generation methods frequently used in the literature. Below are the details of the datasets utilized in our study:

FaceForensics++ : The FaceForensics++ (FF++) benchmark is a widely used dataset that includes deepfake videos generated using multiple state-of-the-art synthesis methods. Specifically, our study leverages 1,000 deepfake videos generated using four synthesis techniques: FaceSwap (FS), DeepFake (DF), Face2Face (F2F), and NeuralTextures (NT). Each video has been processed at various quality levels, and we selected the high-quality subset labeled "C23" for our experiments. This dataset has been extensively cited in the literature (over 2,200 citations) and remains a cornerstone for evaluating deepfake detection techniques.

DeepFakeDetection (DFD): The DeepFakeDetection (DFD) dataset was released by Google AI and is based on an advanced version of the FaceSwap-GAN generation method. It includes 363 real videos recorded by 28 actors and 3,068 corresponding deepfake videos generated using their improved method. Due to its integration into the FaceForensics++ benchmark and its credibility in the research community, this dataset is recognized as a standard for deepfake detection evaluation.

CELEB-DF: The CELEB-DF dataset is an advanced benchmark dataset created using an improved version of the FaceSwap-GAN generation method. It consists of two versions: CELEB-DF-V1 and CELEB-DF-V2. We utilize the larger and higher-quality CELEB-DF-V2, which contains 590 real videos and 5,639 deepfake videos. CELEB-DF-V2 includes challenging and photorealistic deepfakes, making it an excellent benchmark for evaluating detection methods. This dataset is widely accepted in the research community and has been cited over 1,200 times in the literature.

METHODOLOGY

The proposed methodology utilizes a combination of a frozen CLIP vision encoder, multi-scale feature adapters, and learnable class-specific context vectors to enhance deepfake detection.

Input Image Tokens and Transformer Layers

The input image is divided into smaller patches, resulting in a sequence of tokens. These tokens are processed by a series of frozen transformer layers from the CLIP vision encoder. The input to the first transformer layer consists of:

- **Image Class Token** (X_0): A special learnable token initialized to represent the global representation of the image.
- **Image Patch Tokens** ($X_{1\ldots n}$): Derived by dividing the image into n patches, where each patch is flattened and linearly projected into the token embedding space.

Thus, the initial input sequence to the first transformer layer is:

$$X_{input} = [X_0, X_1, X_2, \ldots, X_n]$$

The tokens are processed through the frozen CLIP vision transformer layers.

Multi-Scale Adapters (MSA) for Feature Refinement

To enhance the features of the patch tokens ($X_{1\ldots n}$) during the transformer layer processing, Multi-Scale Adapters (MSAs) are introduced between consecutive transformer layers. MSAs apply additional refinement to the patch tokens using multi-scale convolutional layers.

- **Multi-Scale Convolutions**:

  The patch tokens ($X_{1\ldots n}$) are processed by convolutional layers with varying kernel sizes ($k \in \{1, 3, 5\}$) to extract features at different spatial scales:

  $$F_i = Conv_{k \times k}(X_{1\ldots n})$$

  where

  - $Conv_{1\times1}$: Captures channel-wise interactions (local context).
  - $Conv_{3\times3}$ and $Conv_{5\times5}$: Capture increasingly larger spatial contexts.
- **Feature Aggregation**:

  The outputs from the multi-scale convolutions are concatenated along the channel dimension to create a unified feature map:

$F_{concat} = Concat([F_1, F_2, F_3], axis = channels)$

A 1×1 convolution is applied to reduce the dimensionality and integrate the concatenated features:

$F_{agg} = Conv_{1\times 1}(F_{concat})$

- **Token Refinement**:

  The aggregated features (F_{agg}) are added back to the original patch tokens:

  $X_{1\ldots n}^{refined} = X_{1\ldots n} + F_{agg}$

  The refined tokens are then passed to the next transformer layer, ensuring that both local and global contexts are captured across layers.

Learnable Class-Specific Context Vectors

The architecture includes **two class-specific learnable context vectors** $C \in R^{2 \times n_{ctx} \times d}$, one for each class (real and fake) where n_{ctx} is the number of context tokens and d is the dimension of each token which matches with the dimensionality of CLIP's text encoder.

These context vectors serve as additional semantic representations that aid in distinguishing between the target classes.

- **Enhancing Context Vectors**:

  The context vectors (C) are enhanced by combining them with the refined image patch tokens $X_{1\ldots n}^{refined}$ using multi-head attention:

  - $C_{enhanced} = MultiHeadAttention(C, X_{1\ldots n}^{refined}, X_{1\ldots n}^{refined}) + C$

  where $Query = C, Key = X_{1\ldots n}^{refined}, Value = X_{1\ldots n}^{refined}$. The addition ensures residual learning for stable optimization.

- **Incorporating Class Labels**:

  The token embeddings of the class labels (*real* and *fake*) are concatenated at the end of the respective enhanced context vectors:
  $C_{final} = C_{enhanced} || \, class\ label\ embedding$

- **Processing by CLIP Text Encoder**:

  The final context vectors are processed through CLIP's frozen text encoder to generate text embeddings:

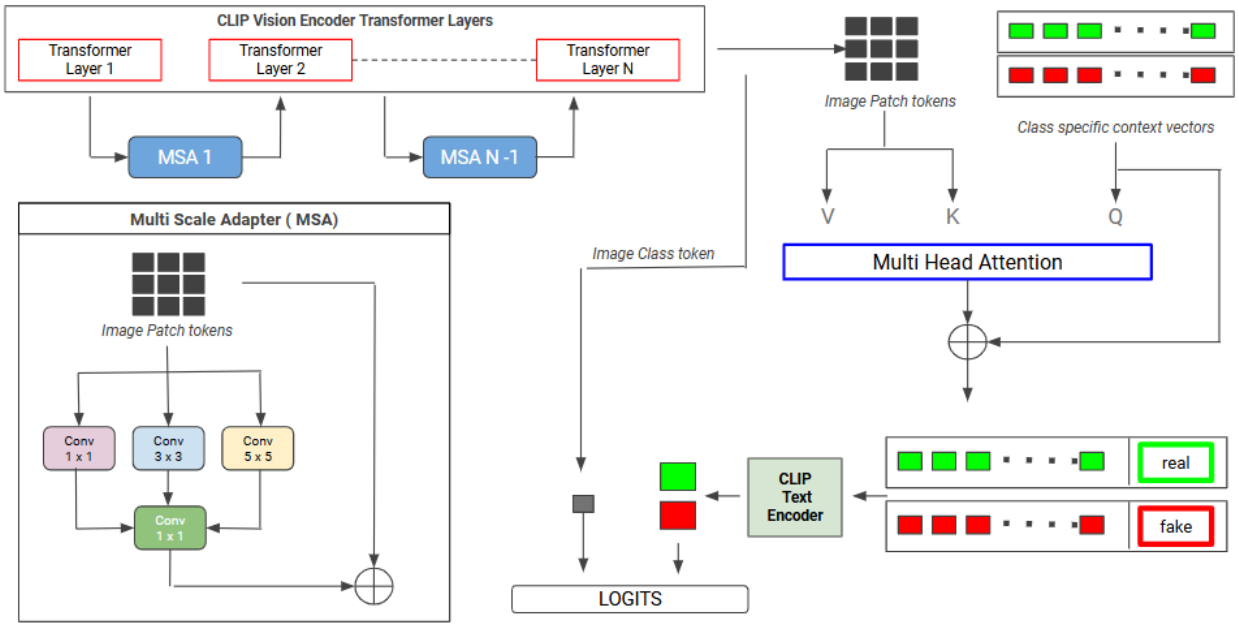
  $E_{text} = TextEncoder(C_{final})$

- **Final Classification**

  The image class token (X_0) is compared to the text embeddings (E_{text}) using cosine similarity. The logits are calculated as:

  $Logits = Softmax(CosineSimilarity(X_0, E_{text}))$

The predicted class corresponds to the highest value in the softmax output. This process enables the model to leverage both image features and contextual semantics for robust deepfake detection.



RESULTS

Compared to existing state-of-the-art approaches, the proposed method effectively balances Seen dataset accuracy with generalization, achieving significant improvements in Unseen dataset performance. Results are demonstrated in Table 1.

MODEL	ACCURACY (%)						Degradation (%)
	Seen	Unseen					
	F2F	CELEB	FS	NT	DF	DFD	
MesoNet	**97.35**	53.13	54.55	52.20	56.00	54.35	43.30
DCT Classifier	92.16	49.95	51.30	55.70	60.80	57.15	37.18
UnivCLIP	89.10	55.93	53.09	58.42	59.20	56.78	32.42
Prompt-tuned CLIP	78.75	54.55	63.15	54.60	67.45	47.75	21.25
Proposed Method	82.15	**64.60**	**66.75**	**60.80**	**75.55**	**63.45**	**15.92**

The proposed method demonstrated the least degradation between Seen and Unseen datasets (15.92%), outperforming MesoNet (43.30%), DCT Classifier (37.18%), UnivCLIP (32.42%), and Prompt-tuned CLIP (21.25%). These results highlight the method's robustness and strong generalization capabilities, making it highly suitable for real-world deepfake detection tasks.