From Pixels to Semantics: A Generalised Approach to Deepfake Detection

Haotong Yu
InfoComm Technology Cluster
Singapore Institute of Technology
Singapore, Singapore
2302705@sit.singaporetech.edu.sg

Shao Ern Toh
InfoComm Technology Cluster
Singapore Institute of Technology
Singapore, Singapore
2302727@sit.singaporetech.edu.sg

Sebastian Nuguid Fernandez
InfoComm Technology Cluster
Singapore Institute of Technology
Singapore, Singapore
2302811@sit.singaporetech.edu.sg

Ashsyahid Bin Hussin
InfoComm Technology Cluster
Singapore Institute of Technology
Singapore, Singapore
2302656@sit.singaporetech.edu.sg

Shaikh Mohamed Irfan
InfoComm Technology Cluster
Singapore Institute of Technology
Singapore, Singapore
2302696@sit.singaporetech.edu.sg

Xiaoxiao Miao[1]
InfoComm Technology Cluster
Singapore Institute of Technology
Singapore, Singapore
xiaoxiao.miao@singaporetech.edu.sg

***Abstract*—The advent of generative artificial intelligence has spurred rapid advancements in media manipulation, resulting in the proliferation of deepfakes. These synthetic media challenge authenticity verification, with current detection methods often failing against novel deepfake types. This paper proposes a generalised approach to deepfake detection using a combination of Vision language model (i.e. CLIP) for extraction of semantic-level information using static prompts, and convolutional neural network (i.e. ResNet101) for extracting hierarchical features from images. Our domainless dataset approach and feature fusion methodology aim to enhance robustness against diverse deepfake techniques. With a detection accuracy of 91.24%, a Detection Cost Function (DCF) of 0.0391, and a Equal Error Rate (EER) of 8.73%. Our model demonstrates adaptability and efficacy in a dynamic media landscape.**

***Index Terms*—deepfake detection, vision-language models, convolutional neural networks, machine learning, media authenticity**

I. Introduction

In recent years, the rapid advancement of generative artificial intelligence (AI) has revolutionized numerous fields, including media creation and manipulation. One significant outcome of this progress is the proliferation of deep-fake images—synthetic media in which a person's likeness is replaced with someone else's or entirely fabricated using machine learning techniques. This technology can produce highly convincing but false visual content, posing challenges for authenticity verification and misinformation management.

The existing solutions for detecting and mitigating deepfakes encounter significant challenges [14]. Many detection algorithms are developed using specific datasets that encompass particular types of manipulations or synthetic content generation techniques. For example, the DeepfakeBench [16] dataset offers a comprehensive collection of domain-specific deepfake data tailored for training models to perform deepfake detection tasks. However, as deepfake technologies rapidly advance, these specialized solutions frequently falter when confronted with new, previously unseen forms of deepfakes. [6] This limitation is further compounded by cross-domain issues: a model trained on one type of data may prove ineffective when applied to another domain. A prime illustration of this problem is a detector calibrated on face-swapped videos potentially failing to identify deepfakes produced by Generative Adversarial Networks (GANs) [4], which fabricate images from scratch rather than modifying pre-existing ones.

Deepfakes can generally be categorized into three main types [7]. **Face-swap**: This involves replacing one person's face with another in a video or image, typically preserving the target's expressions and movements. **Feature enhancement**: Techniques in this category aim to modify or enhance specific features of a subject's appearance, such as smoothing skin or changing hair color. [8]**GAN generation**: Generative methods (i.e. GAN) create entirely new images or videos that did not exist before, sometimes blending characteristics of multiple sources to fabricate a completely new identity. [8]

Our work addresses all three types of deepfakes by training on a domain-less dataset—a collection of data not confined to any particular type of deepfake or source material. This approach aims to develop a solution that is versatile and adaptable to real-world scenarios, where deepfakes can vary widely in form and sophistication.

Central to our approach is the use of Vision-Language Models (VLMs) to provide an extra layer of data insights on top of using a traditional CNN model. VLMs have demonstrated the capability to provide extensive generalization to previously unseen data, thanks to their multi-model architecture that integrates visual and textual information. [5] This allows the model to understand context beyond just the pixels, improving its ability to identify inconsistencies indicative of deepfake alterations. The combination of a domain-less training strategy and VLM's inherent strengths positions our solution as a robust tool against the growing challenge of deepfakes in

[1] This author is the supervisor.

diverse and dynamic environments.

II. Related Work

A. Traditional Approach

Traditional approaches to deepfake detection have primarily focused on leveraging specific characteristics of the manipulated content for identification. For face-swap deepfakes, methods often rely on **contextual anomalies**—inconsistencies that arise from the replacement of one person's face with another's, which may not perfectly match the scene's lighting, pose, or expression [5]. These inconsistencies can be subtle but are detectable by algorithms trained to recognize them.

For instance, Nirkin, Yuval et al. [9] emphasized the differences between the face and the context around the face, exploiting discrepancies in lighting, shadow, and alignment that occur when a face is transplanted into a new environment. Similarly, Yang, Xin et al. [17] introduced the use of face and head positions to identify synthetic images, focusing on how the geometry and positioning of facial features can reveal signs of manipulation.

In addition to contextual analysis, traditional approaches also employ **pixel-wise analysis** for feature enhancement deepfakes, where alterations are more localized and fine-grained. This involves examining the detailed elements of an image, such as skin texture, colour gradients, and other micro-features that might indicate manipulation. By comparing these elements against known patterns of authentic media, it is possible to flag potential deepfakes. In this vein, Ciftci, U.A. et al. [2] extracted biological signals through imaging photoplethysmography (PPG), traditionally used to predict blood pressure levels based on video of the patient's skin. The high sensitivity of PPG to subtle changes in skin colouration made it an effective feature extractor for pixel-wise tasks, enabling the detection of inconsistencies that could signal deepfake manipulations.

The tools underpinning these traditional approaches include Convolutional Neural Networks (CNNs), which excel at identifying spatial hierarchies in images, making them effective for recognizing both contextual and pixel-wise anomalies [11]. Attention models have also been utilized to enhance CNNs' capabilities by focusing on specific regions of an image that are most likely to contain manipulations [18]. Furthermore, data augmentation techniques are employed to increase the diversity of training data, thereby improving a model's ability to generalize to unseen examples [15]. However, despite their strengths, these methods can struggle when faced with the rapidly evolving nature of deepfake generation techniques, especially when dealing with cross-domain challenges.

B. Transformer Approach

In contrast to traditional methods, recent advancements have seen the emergence of Vision-Language Models (VLMs) and multi-modal learning as powerful tools for deepfake detection. VLMs integrate visual and textual information, allowing for a deeper understanding of context and semantics beyond what is possible with purely visual analysis. The transformer architecture underlying VLMs enables these models to capture long-range dependencies and intricate patterns within data, which can be critical for detecting subtle signs of manipulation that might escape traditional models.

The use of transformers has opened up new possibilities for handling multi-modal data, combining visual inputs with textual metadata or descriptions to provide richer insights into the authenticity of media. [13] This approach not only enhances the model's generalization capabilities across different types of deepfakes but also increases its robustness to variations in the source material. By leveraging the strengths of VLMs and multi-modal learning, the latest approaches are better equipped to address the dynamic and diverse landscape of deepfake generation, offering a promising direction for future research and development in this field.

III. Methods

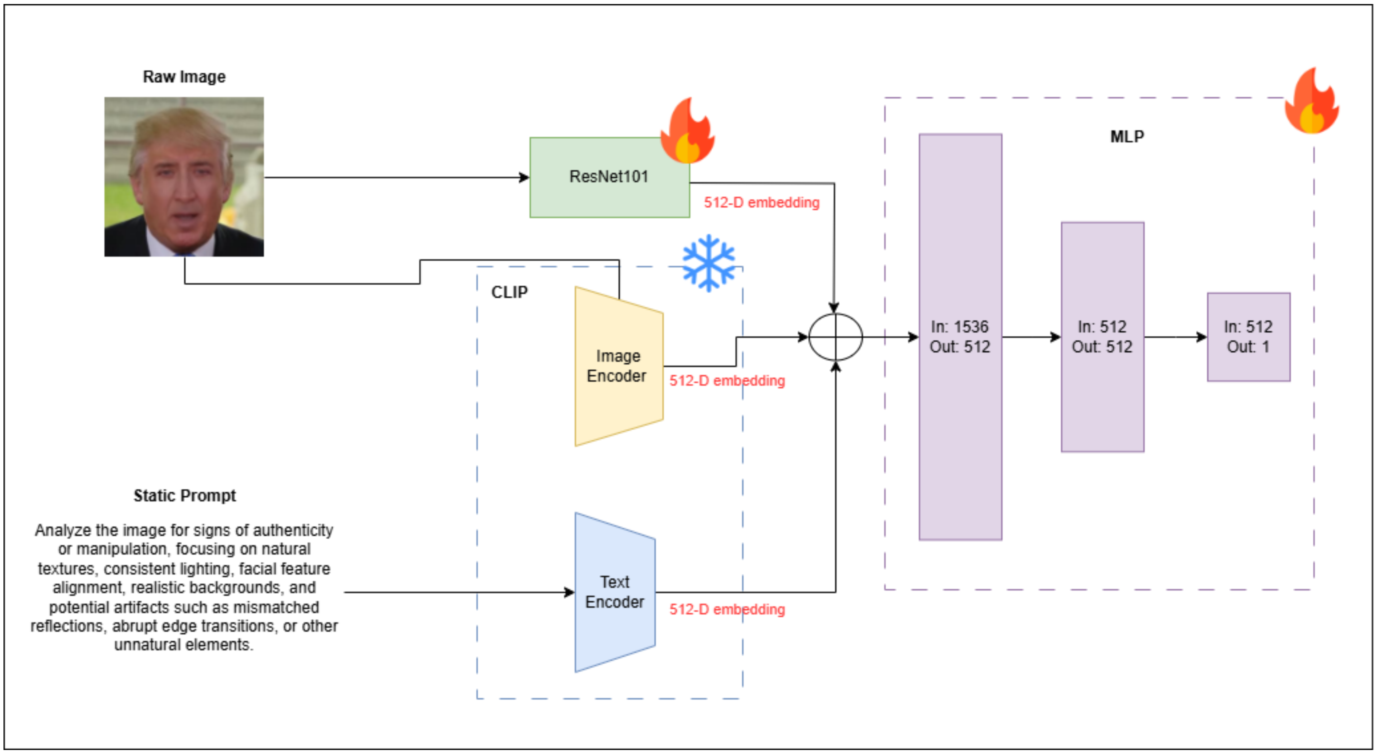
To develop a generalized model capable of detecting diverse types of deepfakes in domainless datasets, we employed a model fusion approach. This design integrates both traditional Convolutional Neural Networks (CNNs) and Vision-Language Models (VLMs), leveraging their complementary strengths to analyze low-level features and high-level semantic information. The overall architecture of our model is depicted in Fig 1.

A. Contrastive Language-Image Pre-Training Model (CLIP)

CLIP was originally designed to measure the similarity between images and text by representing both within a shared feature space, where a higher similarity score indicates a stronger relationship between the image and the text. Building on this concept, we explored the use of the CLIP model from urlhttps://github.com/openai/CLIP as a feature extractor by providing it with a static prompt alongside a deepfake image. This approach enables the model not only to detect pixel-level anomalies but also to identify semantic inconsistencies, such as mismatches between depicted scenes and textual descriptions [10].

To leverage CLIP for deepfake detection, we crafted detailed and structured textual prompts, instructing the model to analyze each image based on critical indicators of manipulation [1]. The designed prompt is as follows: *"Analyze the image for signs of authenticity or manipulation, focusing on natural textures, consistent lighting, facial feature alignment, realistic backgrounds, and potential artifacts such as mismatched reflections, abrupt edge transitions, or other unnatural elements."*

By using these static textual prompts uniformly across all images, we ensured a consistent contextual basis for analysis. The CLIP model outputs embeddings for both the image and the associated prompt, mapping them into the same feature space. This representation allows the model to capture a semantic understanding of the image beyond its visual content, enabling it to identify subtle, contextual anomalies characteristic of deepfakes.



Analyze the image for signs of authenticity or manipulation, focusing on natural textures, consistent lighting, facial feature alignment, realistic backgrounds, and potential artifacts such as mismatched reflections, abrupt edge transitions, or other unnatural elements.

Fig. 1. The proposed model pipeline processes raw images through two separate models: ResNet101 and CLIP. We first finetune the ResNet101 model to extract low-level image features. Then, we use a frozen pretrained CLIP model for extracting semantic-level information using static textual prompts. The extracted features are concatenated and passed through a trainable Multi-Layer Perceptron (MLP) for the final prediction. This architecture integrates contextual and pixel-level insights to enhance detection performance.

B. *Convolutional Model*

For low-level feature extraction, we fine-tuned a ResNet101 model on our curated training dataset. We utilized the ResNet101 implementation from PyTorch's torchvision module, pre-trained on the ImageNet1K dataset. Detailed information about the model's training process can be found at: https://github.com/pytorch/vision/issues/3995#new-recipe.

The ResNet101 architecture is particularly well-suited for capturing intricate spatial patterns, including textures, edges, and localized inconsistencies [3]. These features are critical for identifying manipulations in deepfake images, such as blurred boundaries in face-swaps or unnatural skin textures caused by feature enhancement artifacts [1]. By fine-tuning the model, we adapted it to the unique characteristics of our dataset while leveraging the robust capabilities of its pre-trained weights [12].

C. *Feature Fusion*

The core innovation of our approach lies in the fusion of embeddings derived from the CLIP and ResNet101 models. We concatenated the semantic embeddings from CLIP with the low-level feature embeddings from ResNet101 to create a unified representation of each image. This combined embedding captures both:

- High-level semantic insights, such as contextual consistency and textual relationships.
- Detailed visual features, including spatial and pixel-level anomalies.

The fused embeddings were then passed through a Multi-Layer Perceptron (MLP) with 2 hidden layers, consisting of several fully connected layers. These layers transform the combined feature space into a prediction vector, outputting the likelihood of the image being a deepfake. To ensure robust learning, we employed batch normalization and dropout layers within the MLP to reduce overfitting and improve generalization .

By integrating low-level and high-level features through our model fusion approach and employing robust training and evaluation strategies, our model demonstrates strong potential for accurate and adaptable deepfake detection in dynamic and diverse environments.

IV. Experiments

A. *Data Used*

The dataset used to train the model was obtained from the Deepfake Face In The Wild Competition (DFWild-Cup). It consists of a collection of eight publicly available datasets from DeepfakeBench, divided into training and validation sets containing both real and fake images. The model was trained on 219470 fake images and 42690 real images, and validated on 1524 fake images and 1548 real images.

Simple pre-processing techniques were applied. The images are rescaled to 224×224 pixels and normalized using the ImageNet mean and standard deviation.

V. Environment Setup

The model training was conducted on an NVIDIA RTX A6000 GPU. Our proposed architecture comprises three main components, summarized in Table I. Each component's structural details are elaborated below. multirow

TABLE I
MODEL ARCHITECTURE SPECIFICATIONS

Component	Architecture Details	Params	Output Dim
CLIP Image Encoder	ViT-B/32 Transformer (12-layer, 8-head attention)	151.3M (frozen)	512
CLIP Text Encoder	Transformer (12-layer, 8-head attention)		512
ResNet101	101-layer CNN (33 conv layers, 3 FC layers)	44.5M (trainable)	512
MLP Classifier	3 Linear Layers (1536-512- 512- 1 units, ReLU, Dropout 0.4)	0.5M (trainable)	1
		Total Parameters	195.9M

A. Training

The training setup involves a batch size of 64, using the Adam optimizer with a learning rate of 1e-4. CrossEntropyLoss is employed as the loss function, and training is performed for 30 epochs with Kaiming Initialization and a dropout rate of 0.4 for regularization. The detailed training configurations are presented in Table II.

TABLE II
MODEL TRAINING CONFIGURATIONS

Parameter	Value
Batch Size	64
Learning Rate	1e-4
Optimizer	Adam
Loss Function	CrossEntropyLoss
Number of Epochs	30
Dropout Rate	0.4
Weight Initialization	Kaiming Initialization
Time to Train Per File	0.016970 Second

VI. Evaluation

The evaluation process emphasized measuring the model's effectiveness and robustness using industry-standard metrics:

TABLE III
PERFORMANCE[1]

Scoring	Value
Accuracy	91.24%
Detection Cost Function (DCF)	0.0876
F1 Score	0.9136
Equal Error Rate (EER)	8.73%

VII. Conclusion

This work presents a novel generalised deepfake detection framework combining CNNs and VLMs. By leveraging a domainless dataset and a feature fusion strategy, our model achieves state-of-the-art accuracy and robustness against unseen deepfakes. The integration of textual and visual modalities enables contextual analysis beyond pixel-level features, addressing challenges posed by evolving deepfake technologies. Future work will explore real-time detection capabilities and broader applications to multimedia forensics.

References

[1] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Mesonet: a compact facial video forgery detection network. In *2018 IEEE international workshop on information forensics and security (WIFS)*, pages 1–7. IEEE, 2018.

[2] Umur Aybars Ciftci, Ilke Demir, and Lijun Yin. Fakecatcher: Detection of synthetic portrait videos using biological signals. *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[4] Boquan Li, Jun Sun, and Christopher M Poskitt. How generalizable are deepfake detectors? an empirical study. *arXiv preprint arXiv:2308.04177*, 2023.

[5] Chia Xin Liang, Pu Tian, Caitlyn Heqi Yin, Yao Yua, Wei An-Hou, Li Ming, Tianyang Wang, Ziqian Bi, and Ming Liu. A comprehensive survey and guide to multimodal large language models in vision-language tasks. *arXiv preprint arXiv:2411.06284*, 2024.

[6] Siwei Lyu. Deepfake detection: Current challenges and next steps. In *2020 IEEE international conference on multimedia & expo workshops (ICMEW)*, pages 1–6. IEEE, 2020.

[7] Momina Masood, Mariam Nawaz, Khalid Mahmood Malik, Ali Javed, Aun Irtaza, and Hafiz Malik. Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward. *Applied intelligence*, 53(4):3974–4026, 2023.

[8] Yuval Nirkin, Yosi Keller, and Tal Hassner. Fsgan: Subject agnostic face swapping and reenactment. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7184–7193, 2019.

[9] Yuval Nirkin, Lior Wolf, Yosi Keller, and Tal Hassner. Deepfake detection based on discrepancies between faces and their context. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6111–6121, 2021.

[10] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[11] Mj Alben Richards, E Kaaviya Varshini, N Diviya, P Prakash, P Kasthuri, and A Sasithradevi. Deep fake face detection using convolutional neural networks. In *2023 12th International Conference on Advanced Computing (ICoAC)*, pages 1–5. IEEE, 2023.

[12] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1–11, 2019.

[1] Note: The ablation study involving ResNet101, CLIP-image + ResNet, and CLIP-text-image + ResNet is not included in the current results due to time constraints. However, we plan to incorporate these experiments in future updates to provide a comprehensive analysis.

[13] Junke Wang, Zuxuan Wu, Wenhao Ouyang, Xintong Han, Jingjing Chen, Yu-Gang Jiang, and Ser-Nam Li. M2tr: Multi-modal multi-scale transformers for deepfake detection. In *Proceedings of the 2022 international conference on multimedia retrieval*, pages 615–623, 2022.

[14] Tianyi Wang, Xin Liao, Kam Pui Chow, Xiaodong Lin, and Yinglong Wang. Deepfake detection: a comprehensive study from the reliability perspective. *arXiv preprint arXiv:2211.10881*, 2022.

[15] Zhiyuan Yan, Yuhao Luo, Siwei Lyu, Qingshan Liu, and Baoyuan Wu. Transcending forgery specificity with latent space augmentation for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8984–8994, 2024.

[16] Zhiyuan Yan, Yong Zhang, Xinhang Yuan, Siwei Lyu, and Baoyuan Wu. Deepfakebench: A comprehensive benchmark of deepfake detection. In *Advances in Neural Information Processing Systems*, volume 36, pages 4534–4565. Curran Associates, Inc., 2023.

[17] Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8261–8265. IEEE, 2019.

[18] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. Multi-attentional deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2185–2194, 2021.