Delving into Sequential Patches for Deepfake Detection

# Jiazhi Guan; Hang Zhou; Zhibin Hong et al.

## 2022

We propose a reliable framework to address the practical problems of deepfake detection, which emphasizes on the low-level temporal patterns of sequential patches in the restricted spatial region with a whole-range temporal receptive field using Transformer blocks

# Abstract

Recent advances in face forgery techniques produce nearly visually untraceable deepfake videos, which could be leveraged with malicious intentions. As a result, researchers have been devoted to deepfake detection. Previous studies have identified the importance of local low-level cues and temporal information in pursuit to generalize well across deepfake methods, however, they still suffer from robustness problem against post-processings. In this work, we propose the Local- & Temporal-aware Transformer-based Deepfake Detection (LTTD) framework, which adopts a local-to-global learning protocol with a particular focus on the valuable temporal information within local sequences. Specifically, we propose a Local Sequence Transformer (LST), which models the temporal consistency on sequences of restricted spatial regions, where low-level information is hierarchically enhanced with shallow layers of learned 3D filters. Based on the local temporal embeddings, we then achieve the final classification in a global contrastive way. Extensive experiments on popular datasets validate that our approach effectively spots local forgery cues and achieves state-of-the-art performance.

# Study subjects

**4 datasets**

Our experiments are conducted based on several popular deepfake datasets including FaceForensics++ (FF++) [45], DeepFake Detection Challenge dataset (DFDC) [15], CelebDF-V2 (CelebDF) [35], FaceShift dataset (FaceSh) [30], and DeeperForensics dataset (DeepFo) [24]. **FF++ (HQ) is used as train set and the remaining four datasets are used for generalization evaluation**. FF++ is one of the most widely used dataset in deepfake detection, which contains 1000 real videos collected from Youtube and 4000 fake videos generated by four different forgery methods including Deepfakes [1], FaceSwap [2], Face2Face [53] and NeuralTextures [52]

# Scholarcy Synopsis

Novel attack pattern for DeepFake anti-forensics, namely, the trace removal attack.  
We propose a trace removal network based on an adversarial learning framework involving one generator and multiple discriminators to remove various traces simultaneously.

Deep face forgeries, known as DeepFakes, are raising serious social concerns for information security.  
They focused on an anti-forensics attack against DeepFake detectors.  
The authors presented a novel detector-agnostic attack, called a trace removal attack, that is capable of refining DeepFake images.  
The authors present a novel detector-agnostic attack that is capable of refining DeepFake images by removing all possible DeepFake traces via an one-versus-multiple adversarial learning network.  
The refined Deepfake images are closer to the real images and can bypass arbitrary and even unknown detectors.  
They assessed the efficacy of the trace removal attack against a wide range of state-of-theart detectors.  
  
There were 66000 semantically-closest pairs in the study.

# Findings

Extensive experiments on popular datasets validate that our approach effectively spots local forgery cues and achieves state-of-the-art performance

3) Quantitative experiments show that our approach achieves the state-of-the-art generalizability and robustness

Experiments in [39] show that simple smoothing could impair the performance of more than 20%

# Scholarcy Highlights

* With the development of face forgery methods [1, 25, 26, 66, 28, 23, 37], an enormous amount of fake videos (a.k.a deepfakes) have raised non-neglectable concerns on privacy preservation and information security
* Motivated by the observations above, we propose the Local- & Temporal-aware Transformer-based Deepfake Detection (LTTD) framework, which focuses on patch sequence modeling in deepfake detection with Transformers [17]
* We show the details of the Local Sequence Transformer (LST) in the middle of Fig. 1, which is divided into two parts: the Local Sequence Embedding and the Low-level Enhanced Transformer stages
* We use the best model, “Model 1”, evaluated in their paper with a comparable backbone, “ViT small”, with our method; 4) LTTD w/o LST indicates the model that we replace the proposed LST with commonly used Patch Embedding and Transformer blocks [17]; 5) LTTD w/o Cross-Patch Inconsistency (CPI) is our LTTD framework trained without using LCP I (Eq (13)); 6) LTTD w/o Cross-Patch Aggregation (CPA) represents the model that we replace the CPA module with a simple fully connected classification layer after average pooling the temporal embeddings from all spacial locations
* We propose a reliable framework to address the practical problems of deepfake detection, which emphasizes on the low-level temporal patterns of sequential patches in the restricted spatial region with a whole-range temporal receptive field using Transformer blocks
* Extensive experiments on popular datasets validate that our approach effectively spots local forgery cues and achieves state-of-the-art performance
* Qualitative results further verify that low-level temporal information can lead to stronger generalizability, which could be a guideline for developing better approaches in the future

# Scholarcy Summary

## Introduction

With the development of face forgery methods [1, 25, 26, 66, 28, 23, 37], an enormous amount of fake videos (a.k.a deepfakes) have raised non-neglectable concerns on privacy preservation and information security

To this end, researches have been devoted to the reliable tagging of deepfakes in order to block the propagation of malicious information.

Researches have been devoted to the reliable tagging of deepfakes in order to block the propagation of malicious information

It is still an open problem due to the limited generalization of detection methods and the continuous advances in deepfake creation.

The generalizability of previous methods is typically unsatisfactory when encountering deepfakes generated by unseen techniques

## Methods

Clean CS CC BW GNC GB PX VC Avg/DropFace X-ray [31] 99.8 97.6 88.5 99.1 49.8 63.8 88.6 55.2 77.5/-22.3LipForensics [21] 99.9 99.9 99.6 87.4 73.8 96.1 95.6 95.6 92.6/-7.3 LTTD99.4 98.9 96.4 96.1 82.6 97.5 98.6 95.0 95.0/-4.3Module effects.

1) Xception is the commonly used backbone in deepfake detection; 2) ViT is the most famous vision transformer backbone, where we use the “small” version with embedding dimension of 384; 3) ViViT [8] is a recently published work developed in the self-attention style with spatio-temporal modeling ability for action recognition.

We use the best model, “Model 1”, evaluated in their paper with a comparable backbone, “ViT small”, with our method; 4) LTTD w/o LST indicates the model that we replace the proposed LST with commonly used Patch Embedding and Transformer blocks [17]; 5) LTTD w/o CPI is our LTTD framework trained without using LCP I (Eq (13)); 6) LTTD w/o CPA represents the model that we replace the CPA module with a simple fully connected classification layer after average pooling the temporal embeddings from all spacial locations.

Due to the abundant inductive bias of the convolution, Xception clearly split the four

## Results

Generalizability should be one of the most concerned properties.

It is usually the Achilles’ heel of most deepfake detectors.

Since deepfakes generated by different forgery methods hold different kinds of forgery cues, and overfitting on semantic visual artifacts of the train set can lead to cross-dataset evaluation collapse.

As we show the performance of models trained on FF++ and tested on four unseen datasets in Table.

1, many methods do not perform satisfactorily.

Our approach outperforms all the recently published novel detectors, and achieves a new state of the art of 91.9 AUC% averaged from the four datasets.

Note PatchForensics focuses on local patches, but only narrows the perceptive field by truncating the CNN without considering the relations between patches globally, it shows limited generalizability

## Conclusion

One straight thought might be “just leave the work to self-attention”, since theoretically patches can progressively find the most relative patches at the same spatial room for temporal modeling.

This is nearly impracticable considering both short and long span temporal information is important to our task [65].

We propose to identify the inconsistency by global contrast, because forgery parts should retain heterogeneous temporal patterns compared with the real ones

We achieve this goal through the Cross-Patch Inconsistency loss and the proposed Cross-Patch Aggregation.

In addition to identifying deepfakes, how we can ensure the predictions are credible remains an open problem, hindering the application of all deepfake detectors

# Confirmation of earlier findings

Also, consider possible attacking against the detectors, we evaluate our approach on different types of perturbed videos. **Following previous works [21, 24], we use the script 2 with FF++ and generate seven types of perturbations at five levels**

# Counterpoint to earlier claims

Vaswani et al [55] first propose to using only self-attention, multilayer perceptron, and layer norm to establish a new canonical form, coined Transformer, for natural language processing (NLP). While in **our method, transformer is introduced to achieve patch-sequential temporal learning in a restricted spatial receptive field with a totally different purpose to identify low-level temporal inconsistency**.

Compared with our designs, one straight thought might be “just leave the work to self-attention”, since theoretically patches can progressively find the most relative patches at the same spatial room for temporal modeling. **However, this is nearly impracticable considering both short and long span temporal information is important to our task** [65]

# Limitations

This is a commonly shared limitation that we do not know if the detectors are calibrated well for real-world deployment. In addition to identifying deepfakes, how we can ensure the predictions are credible remains an open problem, hindering the application of all deepfake detectors.