# ADT: Anti-Deepfake Transformer

Ping Wang[†],Kunlin Liu[†],Wenbo Zhou[†],Hang Zhou[‡],Honggu Liu[†],Weiming Zhang[†],Nenghai Yu[†]
[†] University of Science and Technology of China, [‡] Simon Fraser University

icassp 2022 Singapore

4989

## Abstract

Recently almost all the mainstream deepfake detection methods use Convolutional Neural Networks (CNN) as their backbone. However, due to the overreliance on local texture information which is usually determined by forgery methods of training data, these CNN-based methods cannot generalize well to unseen data. In this paper, we propose a novel transformer-based framework to model both global and local information and analyze anomalies of face images. In particular, we design attention leading module, multi-forensics module and variant residual connections for deepfake detection, and leverage token-level contrast loss for more detailed supervision.

**Keywords:** Deepfake Detection, Transformer.

## Introduction

Our key contributions are threefold as below:

1. We propose a novel deepfake detection framework, Anti-Deepfake Transformer (ADT), which pays attention to both global and local information and makes up for the shortcomings of CNN-based methods.
2. We design Attention Leading Module (ALM), Variant Residual Connection (VRC) and Multi-Forensics Module (MFM) to take full advantage of Transformer and introduce contrast loss.
3. Extensive experiments demonstrate that ADT could maintain considerable performance in the intra-dataset evaluation and achieve state-of-the-art in the cross-dataset evaluation in deepfake detection.

## Methods

**Trans-blocks.** Suppose that all the layers have $C$ self-attention heads then the hidden layer features and attention weights can be expressed as follows:

$$\begin{aligned}\mathbf{Z}_l &= [\mathbf{Z}_l^0; \mathbf{Z}_l^1, \mathbf{Z}_l^2, \ldots, \mathbf{Z}_l^N] \\ \mathbf{A}_l &= [[a_l^{0_0}; a_l^{0_1}, .., a_l^{0_N}], .., [a_l^{C_0}; a_l^{C_1}, .., a_l^{C_N}]]\end{aligned} \tag{1}$$

**Attention Leading Module.** To ensure the correspondence between the input token and the attention weight as much as possible, we merge the attention weights of all the previous layers.

$$\mathbf{A}_{final} = \prod_{l=0}^{L-1} \text{Softmax}(\mathbf{A}_l) \tag{2}$$

Then we find the index of the largest attention weight from $\mathbf{A}_{final}$.

$$\mathbf{Z}_{final} = [\mathbf{Z}_{L-1}^0; \mathbf{Z}_{L-1}^{M_1}, \mathbf{Z}_{L-1}^{M_2}, \ldots, \mathbf{Z}_{L-1}^{M_C}] \tag{3}$$

**Variant Residual Connection.** Texture information is an important clue for deepfake detection.we adopt variant residual connections among adjacent trans-blocks.

$$\mathbf{X}_{T_{i+1}} = \mathrm{F}(\mathbf{X}_{T_i}) - \mathbf{X}_{T_i} \tag{4}$$

**Multi-Forensics Module.** We argue that the detection model should not only focus on those high-layer features but also low-layer features, and allow all the features from different levels participate in the final decision.

$$\mathbf{Pred} = \text{Mean}(\mathbf{S}_1, \mathbf{S}_2, \mathbf{S}_3, \mathbf{S}_4). \tag{5}$$

**Training Losses.** Training such a deep transformer network requires strong and detail supervision. We leverage the combination of classification loss(cross-entropy loss) $\mathcal{L}_{cls}$ and token-level contrast loss $\mathcal{L}_{con}$ as training losses. The latter can be described as follows:

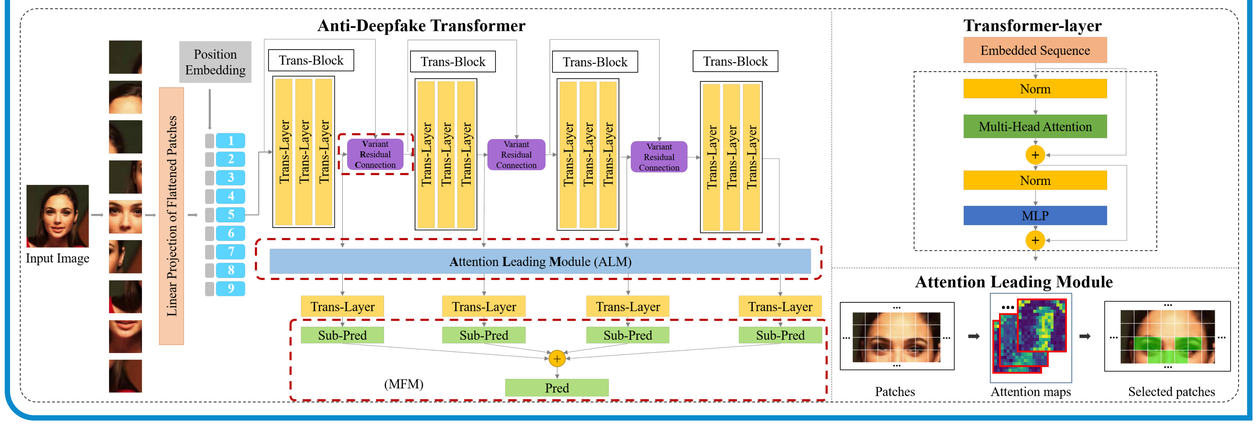
$$\begin{aligned}\mathcal{L}_{con} = \frac{1}{N^2} \sum_i^N \Big( \sum_{j:y_i=y_j}^N (1 - \frac{\mathbf{Z}_i \cdot \mathbf{Z}_j}{\|\mathbf{Z}_i\| \, \|\mathbf{Z}_j\|}) \\ + \sum_{j:y_i \neq y_j}^N \max(\frac{\mathbf{Z}_i \cdot \mathbf{Z}_j}{\|\mathbf{Z}_i\| \, \|\mathbf{Z}_j\|} - \alpha, 0) \Big)\end{aligned} \tag{6}$$

## Acknowledgment

This work was supported in part by the Natural Science Foundation of China under Grant U20B2047, 62072421, 62002334, 62102386 and 62121002, Exploration Fund Project of University of Science and Technology of China under Grant YD3480002001, and by Fundamental Research Funds for the Central Universities under Grant WK2100000011. This work was also partly supported by Shenzhen Key Laboratory of Media Security, Shenzhen University, Shenzhen 518060, China.

## Overview

We show our framework as below. First we split images into small patches and project them into the embedding space. Then we add learnable position embeddings and input them into trans-blocks connected by variant residuals. And then, we apply ALM to select the most valuable tokens from all the tokens output by the final trans-blocks. After that these selected tokens are input into a single transformer layer to get sub-classification. Finally, we merge the four sub-results as the final prediction.




## Results

We compare our framework with state-of-the-art methods in deepfake detection. We train and test the performance of our model on FF++, and further we test the cross-dataset performance on Celeb-DF and other popular datasets to evaluate its transferability.

| Methods | HQ | | LQ | |
|---|---|---|---|---|
| | ACC | AUC | ACC | AUC |
| MesoNet | 83.10 | - | 70.47 | - |
| Face X-Ray | - | 87.35 | - | 61.60 |
| Xception | 92.39 | 94.86 | 80.32 | 81.76 |
| Two-Branch | - | 98.70 | - | 86.59 |
| SPSL | 91.50 | 95.30 | 81.57 | 82.82 |
| $F^3$-Net | 97.52 | 98.10 | 90.43 | 93.30 |
| Multi-attentional | 97.60 | 99.29 | 88.69 | 90.40 |
| M2TR | 98.23 | 99.84 | 92.35 | 94.22 |
| Long-distance | 99.51 | 99.88 | 95.81 | 98.49 |
| Ours | 92.05 | 96.30 | 81.48 | 82.52 |

| Method | FF++ (DF) | Celeb-DF |
|---|---|---|
| MesoNet | 84.70 | 54.80 |
| Xception-c23 | 99.7 | 65.3 |
| Two-Branch | 93.20 | 73.40 |
| SPSL | 96.94 | 76.88 |
| $F^3$-Net | 97.97 | 65.17 |
| Multi-Attention | 99.80 | 67.44 |
| M2TR | 99.50 | 65.70 |
| Long-distance | **99.97** | 70.33 |
| BOLF | - | 78.26 |
| Ours | 98.71 | **84.97** |

## Conclusion

In this paper, we propose a pure transformer-based framework(ADT) for deepfake detection, which aims to expose inconsistency between local and global information. Extensive experiments demonstrate that we achieves the state-of-the-art transferability among almost all the public datasets. And we hope to bring some inspiration.

## Contact Information

**Name** Ping Wang.
**Email** wp123@mail.ustc.edu.cn
**End** Thanks for your attetion.

中国科学技术大学
University of Science and Technology of China