# **Deep Fake Detection Using Wavelet Packets with Vision Transformer (WPT-ViT)**

**Osama M. Rawy and AbdulRahman AlTahhan**

## **Abstract**

## The latest advances in generative algorithms have raised the quality of virtually created images and videos to the point that it has become very difficult to distinguish the real from the generated ones (Deep-Fake). This stimulated hot research to build better models to detect DF. Our paper proposes a new DNN model to detect DF images using a wavelet packet transformer and a vision transformer (WPT-ViT). The study shows that attention could be found between the WPT decompositions of an image even without slicing the image into spatial patches, which is a novel modification to the original VIT model. We showed that by using smaller model sizes and lower GPU and CPU requirements, we can achieve comparable results with previous work in this research area, The model was trained and tested using two datasets, “CIFAKE” and “140k Real and Fake Faces,” which are generated using StyleGAN and Stable Diffusion algorithms.

## **Keywords:** Deep Fake Detection, Wavelet Packets, Vision Transformer.

## **Introduction**

In 2014, Goodfellow et al. introduced the Generative Adversarial Network (GAN), marking the beginning of the generative AI (GAI) era. Since then, researchers have shifted their focus from discriminative learning to generative learning. This wave brought various vision-generative applications to the market, such as Midjourney, Firefly, DALL-E2, and Imagen\cite {bengesi2024advancements}. Such applications were developed using state-of-the-art architectures like GAN, Variational Autoencoders, and Diffusion to generate images and videos with high fidelity and diversity, mimicking real-world photos and videos \cite{raut2024generative}.

Vision-generative technologies have shown high value in several domains. In the entertainment industry, for example, they could generate complete scenes that would otherwise be very risky for actors to perform or prohibitively expensive to produce. In Education, they could bring historical characters to life to talk with students for an immersive learning experience; similarly, the list of positive uses continues in other fields like manufacturing and marketing. However, this ability to produce synthetic content with realistic flavor was termed Deep Fake due to the unfortunate incidents in which these technologies were used to attack people through identity theft, character assassination, and faked pornography. On a larger scale, deep fakes were also used to spread misinformation, fake news, and communal hatred.

According to a report released in April 2021 by Cybernews, deep fake content over the internet doubles every six months, posing a significant threat that needs to be addressed urgently \cite{patel2023deepfake}.

\\For that reason, there has been significant attention in both academic and industrial fields on finding ways to detect deep fakes with high accuracy and performance. For example, Facebook, Microsoft, and Amazon collaborated to launch the Deep Fake Detection Challenge (DFDC) on Kaggle from 2019 to 2020. A survey conducted by (Liang \& Xue, 2024) showed that the number of publications on deep fake detection surpassed the number of publications on deep fake generation in 2022 and 2023 \cite {gong2024contemporary}.

\\This paper introduces a new deepfake detection tool that combines the strengths of wavelet analysis to extract important image features and Vision Transformer (VIT) to create lighter models with lower GPU and CPU requirements than CNN counterparts.

## **Related Work**

Using wavelet transforms to extract the features from signals has shown strength in wide applications; interestingly, a study by (Nadler et al.) showed that the DNNs' color similarity judgments diverge from human color judgments, and a wavelet algorithm provides more coherent color embeddings that better predict human color judgments than all the DNNs examined, ​ They utilized for this purpose a Morlet wavelet to extract textural properties and color information from images [5].

In their paper (Wodajo, et al.,) proposed combining CNN and VIT together to detect both local and global features.

## **Method**

## ***Multiresolution Analysis:***

One key application of Multi-Resolution Analysis (MRA) is to transform a higher-resolution discrete signal into a lower-resolution signal. This has several uses in image processing; the JPEG2000 format uses this concept to provide better compressing performance than the standard JPEG.

The Discrete Wavelet Transform is one example of MRA in which a signal with higher level of details (formally, belongs to a higher-level resolution space  , while is the set of all functions that can be expressed using the basis function ) :

|  |  |
| --- | --- |
|  | (1) |

And then transform it to a lower-level resolution space where  *,* in order to fully represent in the new lower resolution space we need to consider for the error that will result from this transformation by introducing complementary resolution space including all the functions that can be expressed using the basis function

## ***Wavelets:***

Like Fourier analysis, wavelet transform expresses a signal using a set of special functions known as basis.

|  |  |
| --- | --- |
|  | (1) |
|  | (2) |
| *While is the set of all functions that can be expressed using the basis function and is called resolution space.*  *is the set of all functions that can be expressed using the basis function and is called resolution space.* |  |

The basis set starts with a mother function , and other bases will follow by applying a series of time shifts (translations) and scaling (dilations). *Haar* wavelet defined below is one example:

|  |  |
| --- | --- |
|  | (3) |
|  | (4) |
| for & | (5) |
| *While is the time-shift factor, is the scaling factor* |  |

A key difference between wavelets and the sin function that is used as the basis function in Fourier Analysis is that the non-zero part of the wavelet function is limited to a finite time interval (this is the reason behind the name wavelet, which means a small wave); formally the following equation must apply for a function to qualify as a wavelet:

|  |  |
| --- | --- |
|  | (6) |

This last feature is also important to ensure orthogonality between a set of wavelet basis functions if they are arranged so that their nonzero parts don’t overlap.

|  |  |
| --- | --- |
|  | (7) |

## ***Our Classifier:***

In our design we propose passing the image first to wavelet packet transformer that extract the coefficients based on certain wavelet basis function, then use the output of this stage as the input impeding to VIT the encoder stage. These two stages is enough to provide the desired result, however to increase the flexibility of the model we add extra two optional stages in the middle as follows:

*Stage2*: Selects which WPT coefficients should be passed to the model

*Stage3*: Imitates the original VIT design which divide the input image spatially into smaller patches those are used as the input tokens for the encoder stage

|  |
| --- |
|  |
| Figure 1: WPT-VIT Classifier Model |

## **Experiments**

This paper proposes a new method for combining wavelet packet transform with vision transformer to create a binary classifier for image deep-fake detection.

For this purpose, we used two deep faked datasets separately to train and to evaluate the model:

***First: CIFAKE*** which contains 60,000 synthetically generated images using Stable Diffusion version 1.4 and 60,000 real images (collected from CIFAR-10).

The original dimensions of the CIFAKE image is *32 x 32 x 3*

|  |
| --- |
|  |
| Figure 2: CIFAKE Dataset |

***Second: 140k Real and Fake Faces*** which contains 70k REAL faces from the Flickr dataset collected by Nvidia, as well as 70k fake faces sampled from the 1 million FAKE faces (generated by StyleGAN) that was provided by Bojan.

The original dimension of the 140k Real and Fake Faces (hereafter referred to with the letters RVSF) is *256 x 256 x 3*

This dimension was down scaled to *32 x 32 x3* for the model size to fit inside the used GPU.

|  |
| --- |
|  |
| Figure 3: RVSF Dataset |

|  |  |  |  |
| --- | --- | --- | --- |
|  | **REAL** | **FAKE** | **Total** |
| **Training** | 50,000 | 50,000 | 100,000 |
| **Validation** | 10,000 | 10,000 | 20,000 |
| **Testing** | 10,000 | 10,000 | 20,000 |
| **Total** | 70,000 | 70,000 | 140,000 |

Each of the above datasets has been divided into 3 splits (80% for training, 14% for validation and 6% for testing) as follows:

|  |  |
| --- | --- |
|  |  |
| Figure 3: RVSF Dataset | |

Then we performed a series of trials to explore the performance and the flexibility of our model as follows:

## ***Stage1: Use haar, Stage2: Select all, Stage3: Use 1 patch per coefficient:***

Figure 2 shows the curve plot of the Haar function:

|  |  |
| --- | --- |
|  |  |
| *a) Haar scaling function* | *b) Haar wavelet function* |
| Figure 4: Haar function | |

we used this wavelet to extract the WPT coefficients up to a certain level of details, the larger the level the greater number of coefficients and the smaller dimension for each coefficient.

## ***Use CIFAKE and Level 3 of WPT decompositions:***

Sample data point is shown below:

|  |
| --- |
|  |
|  |
| Figure 5: WPT decomposition using haar function and Level 3 |

Then we trained this dataset using single encode with 16 heads for 200 epochs

|  |
| --- |
|  |
|  |
|  |
| Figure 6: CIFAKE training result (Validation Accuracy 90.99%) and Test Recall of 91.92% . |

At the end of the training, we found that an attention has been found between the WPT coefficients of the same image, moreover we found that some coefficients have more influence on the attention matrix than others as shown below:

|  |
| --- |
|  |
|  |
| Figure 7: WPT decomposition using haar function and Level 3 |

## **Conclusion**

This paper proposes a new method for combining wavelet packet transform with vision transformer to

## **References**

1. Percival, D. B., & Walden, A. T. (2000). *Wavelet methods for time series analysis (Vol. 4)*. Cambridge university press.
2. Ryan, Ø., Dahl, G., & Mørken, K. (2014). *Nonlinear optimization Lecture notes for the course MAT-INF2360*. *University of Oslo*.
3. Marcellin, M. W., Gormish, M. J., Bilgin, A., & Boliek, M. P. *(2000, March). An overview of JPEG-2000*. In Proceedings DCC 2000. Data compression conference (pp. 523-541). IEEE.
4. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). *An image is worth 16x16 words: Transformers for image recognition at scale*. arXiv preprint arXiv:2010.11929.
5. Nadler, E. O., Darragh-Ford, E., Desikan, B. S., Conaway, C., Chu, M., Hull, T., & Guilbeault, D. (2023). *Divergences in color perception between deep neural networks and humans.* Cognition, 241, 105621.
6. Wolter, M., Blanke, F., Heese, R., & Garcke, J. (2022). *Wavelet-packets for deepfake image analysis and detection.* Machine Learning, 111(11), 4295-4327.
7. Li, Q., Shen, L., Guo, S., & Lai, Z. (2021*). WaveCNet: Wavelet integrated CNNs to suppress aliasing effect for noise-robust image classification.* IEEE Transactions on Image Processing, 30, 7074-7089.
8. Bird, J. J., & Lotfi, A. (2024). *Cifake: Image classification and explainable identification of ai-generated synthetic images*. IEEE Access.
9. Wodajo, D., & Atnafu, S. (2021). *Deepfake video detection using convolutional vision transformer.* arXiv preprint arXiv:2102.11126.