WaveConViT: Wavelet-Based Convolutional Vision Transformer for Cross-Manipulation Deepfake Video Detection

Mehdi Atamna, Iuliia Tkachenko, Serge Miguet

WaveConViT: Wavelet-Based Convolutional Vision Transformer for Cross-Manipulation Deepfake Video Detection

Mehdi Atamna, Iuliia Tkachenko, and Serge Miguet

Univ Lyon, Univ Lyon 2, CNRS, INSA Lyon, UCBL, Centrale Lyon, LIRIS, UMR5205, F-69676 Bron, France

Abstract. The ease of use and wide availability of high-quality deepfake creation tools raises significant concerns about the reliability and trustworthiness of online content, and makes the task of detecting facial tampering more complicated. As such, the development of effective deepfake detection methods is of utmost importance. In recent years, the facial deepfake detection task took a leap thanks to the development of deep learning-based methods as well as the availability of large datasets of high-quality deepfake videos. Despite the aforementioned methods achieving excellent results when tasked with detecting deepfakes generated using methods seen during training, the cross-manipulation, or generalization, task—where a trained model is exposed to unseen manipulation techniques—is a major challenge which is attracting the attention of the research community. In this paper, we introduce WaveConViT, a novel spatio-temporal architecture for deepfake detection based on Vision Transformers and a two-dimensional discrete wavelet transform. Additionally, we introduce and evaluate a temporal sampling strategy based on frame skipping. We extensively test and benchmark this architecture in the challenging cross-manipulation scenario on the FaceForensics++, Celeb-DF, and DeeperForensics-1.0 datasets, comparing it to a selection of modern, representative Vision Transformer (ViT) and convolutional neural network (CNN) architectures and demonstrating the value of high-frequency features as well as our frame skipping strategy for deepfake detection.

Keywords: Deepfake detection · Video manipulation detection · Discrete wavelet transform · Spatio-temporal features.

1 Introduction

The rapid pace of progress in face swapping and facial reenactment technology has democratized high-visual-quality facial deepfakes, enabling easy access to deepfake creation tools [1, 2]. Although useful in many applications, actors with malicious intent can exploit this technology to spread harmful information, for example by impersonating other people.

In recent years, various methods and datasets [10, 15, 18, 27] have been proposed for deepfake detection. The most effective methods leverage the power of

deep learning for the best possible results [27]. Although many such architectures perform well when evaluated on data derived from the same distribution as the training set, their performance drops dramatically when tasked with detecting deepfakes generated using methods unseen during training (i.e., when exposed to out-of-distribution data).

In this paper, we propose a novel architecture for the detection of deepfake videos consisting of a hybrid backbone leveraging recent advances in convolutional neural network (CNN) and Vision Transformers (ViT) [12], which we have adapted to extract spatio-temporal features. Furthermore, high-frequency feature extraction through a two-dimensional discrete wavelet transform (DWT) block is used, and a cross-attention block for feature enrichment and sharing between the color and high-frequency branches completes the design.

We focus on the cross-manipulation—i.e., generalization—task, where architectures are trained on one dataset and tested on another. This is a more realistic scenario as a trained detector used in the real world is likely to encounter a deepfake generated by a new, unseen tampering method compared to what it was trained on. Specifically, we use three popular, high-quality deepfake video datasets in our experiments: FaceForensics++ [27], DeeperForensics-1.0 [15], and Celeb-DF [18]. These datasets can be broken down into three generations [10] according to factors such as age, scale, and visual quality: first (FaceForensics++), second (Celeb-DF), and third (DeeperForensics-1.0).

A series of experiments using a proposed frame sampling strategy based on frame skipping and comparisons with state-of-the-art CNN- and ViT-based architectures confirm the superior performance of our approach, demonstrating the value of high-frequency and spatio-temporal feature extraction in tackling the generalization task in facial deepfake detection. An ablation study demonstrates the soundness of our approach and validates the usefulness of each component of our architecture.

The rest of the paper is organized as follows. In Section 2, we provide an overview of the state of the art in deepfake detection methods and approaches. In Section 3, we explain our proposed architecture, detailing all of its constituent blocks. In Section 4, we detail our experimental setup including datasets, baseline architectures, data pre-processing, and the training procedure. In Section 5, we present and discuss the results of our proposed architecture and compare its performance with the baseline architectures in both cases—known manipulation techniques and generalization. The ablation study is also discussed in this section. Finally, Section 6 concludes this article.

2 Related Works

Image and video tampering detection has evolved substantially over the years. The earliest detection methods looked at the inconsistency of artifact patterns related to the acquisition chain such as double compression artifacts [5] or irregularities in the sensor-based photo-response non-uniformity [21]. Then, methods

based on handcrafted features for facial deepfake detection appeared, such as methods looking at facial landmark locations [33].

Nowadays, deep learning-based methods achieve excellent performance on modern facial deepfake datasets [14, 22, 26, 27, 32]. Extensive overviews of the different techniques in the field of deepfake detection are provided in [4, 30].

When it comes to deep learning-based methods, various approaches are explored in the literature. For the extraction of visual features from images, CNN architectures have been extensively used and benchmarked for binary image classification [27]. More recently, ViT-based architectures have been proposed for image classification. In [11], the authors propose Identity Consistency Transformer (ICT), a ViT-based architecture that focuses on the consistency of the identity of the subject and achieves promising results, although it is limited mainly to the detection of face swapping. In [31], the authors propose CViT, an architecture which increases the inductive bias of a Transformer by combining it with a convolution-based feature extraction block. In [7], a similar philosophy is pursued, with two CNN feature extractors covering small and large receptive fields being combined with a ViT to form the proposed Convolutional Cross ViT.

Some architectures leverage the temporal dimension in order to classify image *sequences*. In [14, 32], CNNs are combined with a long short-term memory (LSTM) network for the extraction of spatio-temporal features for deepfake detection. In [26], various 3D CNN architectures combined with attention mechanisms are benchmarked in various scenarios. The intuition with these approaches is to exploit the temporal inconsistencies between successive frames in a deepfake video since these are typically generated on a frame-by-frame basis.

In addition to learning features from RGB images, some works exploit frequency-based features to further enrich the learning process. In [25], an architecture is proposed which aims at learning frequency-aware forgery patterns by decomposing the input image into frequency-based features maps. This is accomplished by applying a two-dimensional discrete cosine transform, masking the output with a number of masks (where each mask filters out certain frequency bands), and finally applying the inverse transform to reconstruct the input image in the spatial domain while targeting a specific frequency band for each obtained map. A second stream aims to detect local abnormal frequency distributions by computing image statistics. Other works instead focus on using spatial rich model (SRM) [13] high-frequency filter blocks to extract frequency-based features. In [22], the authors use these filters at multiple levels in their proposed architecture, while in [3], these filters are shown to be efficient and effective at improving results on the generalization task. In [24], the authors propose a Transformer-like network for deepfake image detection which uses a discrete wavelet transform for high-frequency feature extraction.

In [16], the authors approach the deepfake detection problem from a metric-learning perspective, proposing a loss function that aims to minimize the distance between representations of natural faces and the center point without restricting the intra-class compactness of manipulated faces. The intuition is that since

different manipulation techniques are grouped into one class, not imposing a compactness constraint leads to more discriminative features.

Other works focus more on the data side of the pipeline. In [17], the authors use randomly deformable soft facial masks to swap faces between subjects in real videos, thus generating their own fake faces and augmenting the training set. In [28], a method for generating extra training data from a single base image is introduced. To generate new images, a mask generator module uses facial landmarks to produce a geometrically deformed mask. This mask, along with its inverse, is then used to blend the source and target images, which are themselves augmented versions of the base image.

3 Method

Our proposed architecture relies on three fundamental building blocks: two spatio-temporal processing streams (one for RGB features, the other for high-frequency features) and a cross-attention module to allow feature fusion and sharing by linking the two branches. Fig. 1 illustrates the full proposed architecture.

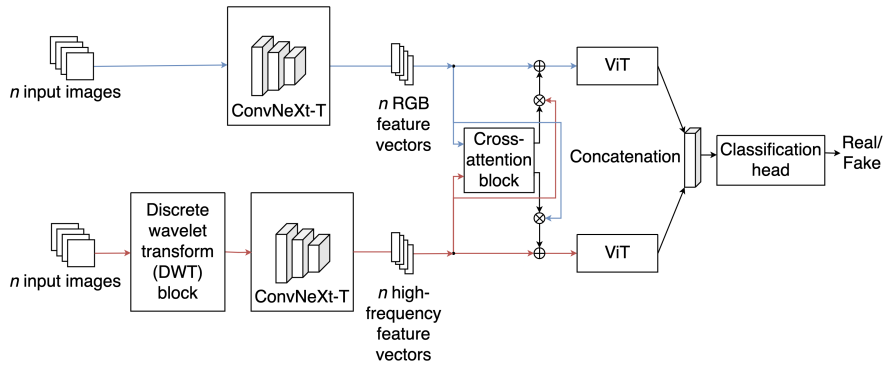

Fig. 1: The proposed WaveConViT architecture. $\oplus$ indicates element-wise addition and $\otimes$ indicates matrix multiplication.

3.1 Spatio-Temporal RGB Feature Extraction

Spatio-temporal feature extraction is accomplished using a hybrid architecture consisting of an ImageNet-1K-initialized ConvNeXt-T [19] convolutional backbone combined with a standard ViT [12]. First, the linear classification layer of ConvNeXt-T is removed. This allows us to obtain, for a temporal sequence of n input images, n feature vectors, the spatial features of each image being summarized in a single vector. These n feature vectors are then fed as input to a ViT

with two layers, four attention heads, and no classification head. In the original ViT, which was used for image classification, the input image was partitioned into patches, each of which was projected and made into a vector-shaped token. In our architecture, however, each input token to the ViT corresponds to a whole image from the input sequence (instead of an image patch). This modification allows our CNN-ViT hybrid architecture to process sequences of input images in a spatio-temporal manner. The final feature vector output by the ViT thus summarizes all of the information from our input image sequence. The upper branch in Fig. 1 illustrates this RGB feature extraction stream.

3.2 High-Frequency Feature Extraction

The synthetic manipulation of a region in an image (i.e., the face) leads to the appearance of local inconsistencies in the modified image, such as improperly masked blending boundaries in face-swapped images. As such, to better highlight these subtle manipulation artifacts and supplement the RGB data, we make use of a high-frequency feature extraction stream which relies on a two-dimensional DWT block. This block is inserted at the beginning of the stream, which is otherwise identical to the RGB feature extraction stream (see the lower branch in Fig. 1).

The DWT block used relies on a single-level two-dimensional transform which uses the discrete Haar wavelet. Two of the simpler wavelets from the Daubechies family of wavelets [8], db1 and db2, were also evaluated in the early stages of experimentation; however, we settled on the Haar wavelet as no appreciable difference in performance was observed. The high-frequency filter f_h and low-frequency filter f_l components of this wavelet are given in Eq. 1 and Eq. 2, respectively:

$$f_h = \frac{1}{\sqrt{2}}\{-1, 1\}, \tag{1}$$

$$f_l = \frac{1}{\sqrt{2}}\{1, 1\}. \tag{2}$$

First, the input color image is convolved using a stride of two with f_h and f_l, which generate respectively H and L, halving the number of columns. Second, these two maps are convolved with the transposes of f_h and f_l, yielding a total of four frequency-based decomposition maps of the input image with half its spatial resolution:

- LL (L convolved with f_l^T),
- HL (H convolved with f_l^T),
- LH (L convolved with f_h^T),
- HH (H convolved with f_h^T).

To avoid further weakening the already-subtle manipulation traces, we do not double the spatial dimensions of the input images through interpolation

before applying the wavelet transform. The first convolution layer of ConvNeXt-T in this branch is also modified to accommodate the fourfold increase in the number of channels resulting from the stacking of the decomposition maps. In our implementation, the LL low-frequency component is preserved as we did not observe any improvement in performance from discarding it. Fig. 2 illustrates how this approach can highlight facial tampering artifacts. In this example, a strong response can be observed on the high-frequency maps on the edges of the cheeks/chin (*Face2Face*) and the eyes (*FaceSwap*).

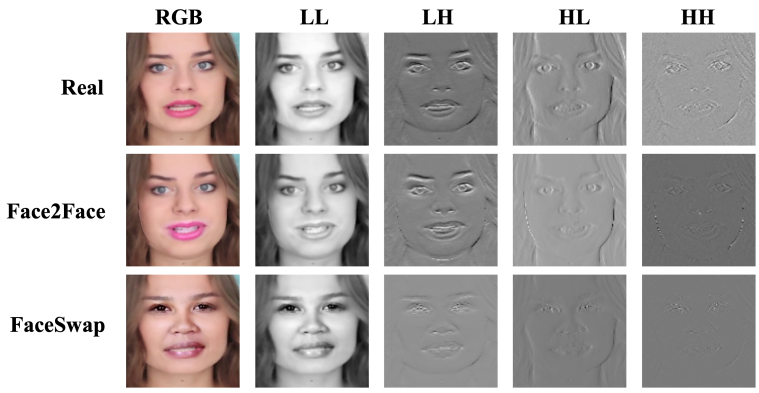

Fig. 2: Comparison of the resulting frequency maps from applying the 2D DWT.

3.3 Cross-Attention Module

In order to enhance interaction between the two main branches, a cross-attention module similar to the one proposed in [25] is applied after the first stage of the feature extraction process (i.e., to the output of the two ConvNeXt-T extractors). The incoming feature vectors are first passed through a dimension-preserving linear layer before the cross-attention scores are computed. Finally, the weighted feature vectors from each branch are added element-wise to the opposite branch to complete the feature enrichment process.

3.4 Classification Head

The output vectors of the two ViTs are concatenated into a single vector of dimension $d = 2048 \times 2$ which is then passed into a classification head. This head consists of two fully connected layers separated by a ReLU nonlinear function. The first linear layer reduces the feature size to 2048 and the second produces the scores for binary classification.

4 Experimental Setup

4.1 Datasets

In our experiments, we use three state-of-the-art facial deepfake video datasets:

(i) FaceForensics++ (FF++) [27]: Contains 1,000 real videos and 5,000 fake videos made by editing the real videos using five different manipulation techniques. These manipulation techniques perform face swapping (*DeepFakes* (DF), *FaceShifter* (FSh), *FaceSwap* (FS)) and facial reenactment (*Face2Face* (F2F), *NeuralTextures* (NT)). We use the H.264, lightly compressed (HQ) version of the dataset.

(ii) Celeb-DF [18]: Contains 5,639 fake and 590 real videos of various celebrities, the type of tampering implemented being face swapping. Videos are compressed to the MPEG4.0 format.

(iii) DeeperForensics-1.0 (DF-1.0) [15]: We use the standard (*std*) set which contains 1,000 fake videos obtained from the same non-manipulated videos as FaceForensics++ using the same HQ compression scheme. The fake videos in this dataset are obtained by face swapping using a different method compared to FaceForensics++.

4.2 Data Pre-Processing

For each video, the first 120 frames are extracted, and the state-of-the-art MTCNN [34] face detector is used to crop the subject's face in each frame. We manually verify the correct detection of the subject's face. The resulting crop is enlarged by 20% in both height and width to guarantee the presence of both manipulated and authentic regions then resized to 224×224 using bilinear interpolation. We use the official training, validation, and test splits recommended by the authors for FF++ and Celeb-DF. Since DF-1.0 (*std*) is generated from the same source videos as FF++, the same FF++ video IDs are used to split DF-1.0 into the various sets.

Image Sequence Classification and Frame Sampling Strategies In order to study the impact of sequence length on classification performance, we evaluate three different length values: each sequence of 10, 20, or 60 successive frames is taken as a data point for the image sequence classification task.

In addition to sampling every frame successively (i.e., a skip of zero), we evaluate other sampling schemes: for sequences of lengths 10 and 20, we test skips of one, two, and three frames before sampling again. For sequences of length 60, we test a skip of one[1]. In order to exploit all available images in a given video, we also shift the sampling window when using frame skipping. The intuition for implementing this frame skipping strategy is that two successive frames in a video may be too similar for the learning of meaningful temporal

[1]Since our videos consist of 120 frames, one is the maximum possible skip for sequences of 60 frames.

features. It is worth noting that different frame rates between videos can result in differences in the elapsed time between two successive samples.

To illustrate this sampling strategy, consider the following example: Let us represent an arbitrary video with $N = 120$ frames as $\{I_i\}_{i=0}^{i=N-1}$. If we were to choose a sequence length of 10 and a frame skip of zero (referred to in Section 5 as *10f*), the resulting set of image sequences we would obtain is described by the following equation:

$$S_{10f} = \big\{\{I_0, I_1, \ldots, I_9\}, \{I_{10}, I_{11}, \ldots, I_{19}\}, \ldots, \{I_{110}, I_{111}, \ldots, I_{119}\}\big\}. \quad (3)$$

If we were to choose the same sequence length but a frame skip of one instead of zero (referred to in Section 5 as *10f, s = 1*), then the resulting set is described by

$$S_{10f,\,s=1} = \big\{\{I_0, I_2, \ldots, I_{18}\}, \{I_{20}, I_{22}, \ldots, I_{38}\}, \ldots, \{I_{100}, I_{102}, \ldots, I_{118}\}, \{I_1, I_3, \ldots, I_{19}\}, \{I_{21}, I_{23}, \ldots, I_{39}\}, \ldots, \{I_{101}, I_{103}, \ldots, I_{119}\}\big\}. \quad (4)$$

Note that in both cases, we would obtain the same number of sequences: $|S_{10f}| = |S_{10f,\,s=1}| = 12$ per video of 120 frames.

4.3 Baseline Architectures

In this work, we use a set of popular baseline architectures to compare and benchmark our proposed method against. For image classification, we use XceptionNet [6], a popular CNN architecture for image classification which is widely used for benchmarking purposes in deepfake detection [27]. We also use the T (tiny) and B (base) versions of ConvNeXt [19], a more modern CNN architecture which achieves excellent results in image classification on ImageNet [9]. All CNN architectures are initialized with pre-trained weights on ImageNet-1K before training on our datasets.

Additionally, we use three Transformer-based deepfake detection architectures which achieve strong results and for which code is publicly available: CViT [31], Efficient ViT [7], and Convolutional Cross ViT [7]. For both Efficient ViT and Convolutional Cross ViT, we use an EfficientNet-B0 [29] that is pre-trained on ImageNet as a feature extractor, as per the original paper.

For image sequence classification, we use the same spatio-temporal architecture in the RGB domain as [3] due to the ease of reproduction of this work. This architecture combines XceptionNet with a single-layer LSTM.

4.4 Training Procedure

We use binary cross-entropy with appropriate class weights to account for the imbalance between real and fake data. For all image classification architectures, a batch size of 32 is used. For the spatio-temporal architectures, due to memory

constraints, we use batch sizes of 8, 4, and 2 with sequences of 10, 20, and 60 frames, respectively. We have found the AdamW [20] optimizer with a learning rate of 10^{-5} and 8 epochs to work well for all tested architectures.

Classification accuracy and the area under the ROC curve (AUC) are used as performance metrics. During training, evaluation on the validation set is carried out periodically four times per epoch, and the model with the highest average of accuracy and AUC is kept at the end of training. Performance scores are computed over all data points in the test set of the target dataset (i.e., we do not compute nor average per-video scores).

5 Results

Before comparing the proposed WaveConViT with the baselines, we present in Table 1 the complete results of WaveConViT in both the standard setting (training and testing on FF++) and the cross-manipulation scenario (training on FF++ and testing on either Celeb-DF or DF-1.0).

	Training & testing on FF++		Generalization			
			Celeb-DF		DF-1.0	
	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC
WaveConViT 10f	96.57	95.59	66.83	60.88	72.50	43.45
WaveConViT 10f, s = 1	94.82	86.74	**71.33**	**73.29**	65.80	45.12
WaveConViT 10f, s = 2	**97.12**	96.46	67.12	59.81	<u>81.25</u>	58.42
WaveConViT 10f, s = 3	96.79	97.74	<u>69.87</u>	<u>67.75</u>	78.21	59.75
WaveConViT 20f	96.83	96.61	67.17	55.14	73.99	45.32
WaveConViT 20f, s = 1	96.63	96.89	67.48	61.49	77.32	52.16
WaveConViT 20f, s = 2	96.37	**98.74**	64.63	52.60	**84.40**	**79.37**
WaveConViT 20f, s = 3	<u>96.96</u>	95.91	63.51	53.09	77.95	48.14
WaveConViT 60f	<u>96.96</u>	96.80	67.15	57.21	77.86	51.16
WaveConViT 60f, s = 1	96.37	<u>97.87</u>	66.34	55.80	79.46	<u>60.97</u>

Table 1: Full classification results of the proposed WaveConViT for all configurations. For each dataset, the best result is indicated in bold, while the second-best is underlined.

Most sequence length and frame skip size configurations achieve comparable and strong performance scores when training and testing is done on FF++. Note, however, how the shortest possible length with a small skip achieves the best results in the generalization test on Celeb-DF while a longer sequence length gives

the best results on DF-1.0. As such, the best configuration is heavily dependent on the target dataset for cross-manipulation detection; however, the *10f, s = 3* configuration achieves a good balance between both datasets.

Another observation is that the accuracy, on average, tends to be higher on DF-1.0 compared to Celeb-DF. This is to be expected since DF-1.0 shares a stronger similarity to the training set since it was created from the same source videos as FF++.

5.1 Known Manipulation Techniques

Table 2 shows the best results for the spatial and spatio-temporal architectures when training and testing on FF++. For readability purposes, only the sequence length and skip size configuration which achieves the highest accuracy for each spatio-temporal architecture is shown.[2]

	Accuracy	AUC
XceptionNet [6]	96.22	87.64
ConvNeXt-T [19]	<u>96.79</u>	80.70
ConvNeXt-B [19]	96.37	88.74
CViT [31]	91.54	85.19
Efficient ViT [7]	91.61	90.24
Convolutional Cross ViT [7]	95.14	<u>94.23</u>
XceptionNet + LSTM 20f, s = 2 [3]	95.81	89.79
WaveConViT 10f, s = 2 (ours)	**97.12**	**96.46**

Table 2: Classification results when training and testing on FF++. The double horizontal line denotes the separation between image and image sequence classification. The best result is indicated in bold while the second-best is underlined.

Our proposed WaveConViT outperforms all baselines both in accuracy and AUC with the *10f, s = 2* configuration. It improves on ConvNeXt-T, which achieves the highest accuracy among the baselines by 0.33% and its AUC is 2.23% higher than second-placed Convolutional Cross ViT. Note that relatively shorter sequences of 10 or 20 frames with frame skipping yield the best results for the spatio-temporal architectures.

[2]The full results and architecture code are available at *gitlab.liris.cnrs.fr/unmask-deepfake-videos/waveconvit*.

5.2 Generalization Performance

Table 3 shows the best results for all architectures in the cross-manipulation (i.e., generalization) scenario. Similarly to Table 2, only the sequence length and skip size configuration with the highest accuracy on either dataset is shown for each spatio-temporal architecture.

FaceForensics++ was chosen for training because it contains fake videos generated using five different techniques. This results in a richer set of tampering clues to learn from compared to Celeb-DF and DF-1.0, both of which use a single tampering technique. As shown in [3], frequency-aware deepfake detection works better when learning occurs on a diverse set of tampering artifacts.

Once again, WaveConViT outperforms all baselines, achieving the best generalization performance overall. The *10f, s = 1* configuration achieves the best performance on Celeb-DF while the *20f, s = 2* configuration is the best on DF-1.0, where WaveConViT significantly outperforms the baselines. Overall, the spatio-temporal architectures outperform the image-only approaches in generalization, showing the value of exploiting the temporal dimension in deepfake video detection.

One important aspect to note with sequence length is that there is a trade-off between the richness of temporal features in longer sequences and the amount of data available for training. Indeed, longer sequences contain more temporal information—potentially enabling the learning of richer and more representative features—but yield fewer total data points for training/testing since each sequence covers a larger part of the total length of the video. Table 4 illustrates this, showing the total number of data points depending on the length of the sequence and the value of the frame skip.

This trade-off may help explain why WaveConViT performs better with relatively shorter sequences, as Vision Transformers are notorious for performing best when trained on larger datasets [12].

5.3 Ablation Study

Table 5 presents the performance in both the known manipulation and cross-dataset evaluation scenarios when isolating specific components of the WaveConViT architecture: RGB branch only, DWT branch only, both branches without the cross-attention module, and the full architecture.

Additionally, we compare our DWT-based high-frequency feature extraction method to SRM [13], another approach which has been used recently for deepfake detection with good results [3, 23] as well as for more general image tampering detection [35]. Specifically, we replace DWT in WaveConViT with a block which uses the same group of SRM kernels as [35] to extract high-frequency maps.

We can observe that the RGB branch outperforms the high-frequency branch when both are evaluated on their own. This reinforces the findings in [3], where high-frequency filtering was shown to suppress meaningful temporal manipulation artifacts such as color inconsistency between successive frames in a deepfake video or flickering artifacts.

	Celeb-DF		DF-1.0	
	Accuracy	AUC	Accuracy	AUC
XceptionNet [6]	68.40	63.69	72.10	30.19
ConvNeXt-T [19]	69.78	59.78	62.58	28.55
ConvNeXt-B [19]	<u>70.74</u>	65.63	71.52	27.74
CViT [31]	69.25	71.00	63.94	52.43
Efficient ViT [7]	66.58	62.15	63.46	53.96
Convolutional Cross ViT [7]	69.45	64.34	67.63	43.97
XceptionNet + LSTM 20f, s = 2 [3]	59.53	50.26	<u>81.49</u>	50.21
XceptionNet + LSTM 60f, s = 1 [3]	67.51	<u>71.03</u>	55.00	<u>64.26</u>
WaveConViT 10f, s = 1 (ours)	**71.33**	**73.29**	65.80	45.12
WaveConViT 20f, s = 2 (ours)	64.63	52.60	**84.40**	**79.37**

Table 3: Classification results when training on FF++ and testing on Celeb-DF and DF-1.0 (i.e., generalization). The double horizontal line denotes the separation between image and image sequence classification. The best result on each dataset is indicated in bold while the second-best is underlined.

Sequence length and frame skip combination	10f				20f				60f	
	no skip	s = 1	s = 2	s = 3	no skip	s = 1	s = 2	s = 3	no skip	s = 1
# of data points (thousands)	72	72	72	72	36	36	36	24	12	12

Table 4: Number of data points (image sequences) available in our FF++ dataset depending on the sequence length and frame skip combination.

Generalization performance increases when both branches are used together and, finally, the best results are achieved when adding the cross-attention module, which raises generalization accuracy by 0.73% and 1.6% on Celeb-DF and DF-1.0, respectively. These tests thus confirm the validity of the design choices underpinning the WaveConViT architecture.

Finally, the comparison in Table 5 shows that the DWT approach using the Haar wavelet allows WaveConViT to learn more generalizable tampering artifacts, resulting in superior cross-dataset performance compared to SRM.

6 Conclusion

In this paper, we propose WaveConViT, a novel spatio-temporal architecture for deepfake video detection which relies on a hybrid backbone consisting of a state-

Components				Training & testing on FF++		Generalization			
						Celeb-DF		DF-1.0	
RGB	DWT	SRM	Cross-attention module	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC
✓				**97.26**	97.07	69.74	64.59	80.36	59.07
	✓			85.97	78.23	67.71	65.11	57.92	54.15
✓	✓			96.85	92.47	70.60	72.79	82.80	58.93
✓	✓		✓	97.12	96.46	**71.33**	**73.29**	**84.40**	**79.37**
✓		✓	✓	96.87	**97.38**	69.50	67.30	84.29	71.09

Table 5: Influence of each component of WaveConViT on classification performance. SRM [3,13], another high-frequency feature extraction method, is shown for comparison against our DWT approach.

of-the-art convolutional neural network and a Vision Transformer, and learns from color images as well as high-frequency data, which is extracted through a two-dimensional discrete wavelet transform. The two constituent streams of WaveConViT—the color and high-frequency ones—can also act as standalone architectures which are also discussed in our study.

Additionally, we introduce a frame-skip-based sampling strategy for creating image sequences, intuiting that successive frames in a video may be too similar for the learning of meaningful temporal features.

Through a series of experiments on the FaceForensics++, Celeb-DF, and DeeperForensics-1.0 datasets—with a particular focus on the challenging cross-manipulation scenario—we test the validity and superiority of our architecture as well as the effectiveness of the aforementioned frame sampling strategy. Specifically, we compare our proposed architecture to a group of state-of-the-art architectures consisting of Transformer-based models for deepfake detection and one of the best-performing convolutional architectures for image classification. Results show WaveConViT outperforming all baselines, demonstrating the potency of temporal and high-frequency features in tackling the challenging deepfake generalization task, as well as the effectiveness of frame skipping strategies.

Future work will focus on the possibility of incorporating learning into high-frequency feature extraction, and the viability of adapting video Vision Transformers to the deepfake detection task.

References

1. DeepFakes. `https://github.com/deepfakes/faceswap`, Accessed: March 2024
2. FaceSwap. `https://github.com/MarekKowalski/FaceSwap/`, Accessed: March 2024

3. Atamna, M., Tkachenko, I., Miguet, S.: Improving Generalization in Facial Manipulation Detection Using Image Noise Residuals and Temporal Features. In: 2023 IEEE International Conference on Image Processing (ICIP). pp. 3424–3428 (2023)
4. Bhagtani, K., Yadav, A.K.S., Bartusiak, E.R., Xiang, Z., Shao, R., Baireddy, S., Delp, E.J.: An Overview of Recent Work in Multimedia Forensics. In: 2022 IEEE 5th International Conference on Multimedia Information Processing and Retrieval (MIPR). pp. 324–329 (2022)
5. Chen, Y.L., Hsu, C.T.: Detecting Recompression of JPEG Images via Periodicity Analysis of Compression Artifacts for Tampering Detection. IEEE Transactions on Information Forensics and Security **6**(2), 396–406 (2011)
6. Chollet, F.: Xception: Deep Learning with Depthwise Separable Convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
7. Coccomini, D.A., Messina, N., Gennaro, C., Falchi, F.: Combining EfficientNet and Vision Transformers for Video Deepfake Detection. In: Image Analysis and Processing – ICIAP 2022. pp. 219–229. Springer International Publishing, 2022
8. Daubechies, I.: The wavelet transform, time-frequency localization and signal analysis. IEEE Transactions on Information Theory **36**(5), 961–1005 (1990)
9. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255 (2009)
10. Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., Canton-Ferrer, C.: The DeepFake Detection Challenge Dataset. ArXiv **abs/2006.07397** (2020)
11. Dong, X., Bao, J., Chen, D., Zhang, T., Zhang, W., Yu, N., Chen, D., Wen, F., Guo, B.: Protecting Celebrities from DeepFake with Identity Consistency Transformer. arXiv preprint arXiv:2203.01318 (2022)
12. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In: International Conference on Learning Representations (2021)
13. Fridrich, J., Kodovsky, J.: Rich Models for Steganalysis of Digital Images. IEEE Transactions on Information Forensics and Security **7**(3), 868–882 (2012)
14. Güera, D., Delp, E.J.: Deepfake Video Detection Using Recurrent Neural Networks. In: 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). pp. 1–6 (2018)
15. Jiang, L., Li, R., Wu, W., Qian, C., Loy, C.C.: DeeperForensics-1.0: A Large-Scale Dataset for Real-World Face Forgery Detection. In: CVPR (2020)
16. Li, J., Xie, H., Li, J., Wang, Z., Zhang, Y.: Frequency-Aware Discriminative Feature Learning Supervised by Single-Center Loss for Face Forgery Detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6458–6467 (June 2021)
17. Li, L., Bao, J., Zhang, T., Yang, H., Chen, D., Wen, F., Guo, B.: Face X-Ray for More General Face Forgery Detection. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
18. Li, Y., Yang, X., Sun, P., Qi, H., Lyu, S.: Celeb-DF: A Large-Scale Challenging Dataset for Deepfake Forensics. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3204–3213 (2020)
19. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A ConvNet for the 2020s. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)

20. Loshchilov, I., Hutter, F.: Decoupled Weight Decay Regularization. In: International Conference on Learning Representations (2019)
21. Lukáš, J., Fridrich, J., Goljan, M.: Detecting digital image forgeries using sensor pattern noise. In: III, E.J.D., Wong, P.W. (eds.) Security, Steganography, and Watermarking of Multimedia Contents VIII. vol. 6072, p. 60720Y. International Society for Optics and Photonics, SPIE (2006)
22. Luo, Y., Zhang, Y., Yan, J., Liu, W.: Generalizing Face Forgery Detection with High-frequency Features. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16312–16321. IEEE Computer Society, Los Alamitos, CA, USA (jun 2021)
23. Luo, Y., Zhang, Y., Yan, J., Liu, W.: Generalizing Face Forgery Detection With High-Frequency Features. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16317–16326 (June 2021)
24. Miao, C., Tan, Z., Chu, Q., Liu, H., Hu, H., Yu, N.: F2Trans: High-Frequency Fine-Grained Transformer for Face Forgery Detection. IEEE Transactions on Information Forensics and Security **18**, 1039–1051 (2023)
25. Qian, Y., Yin, G., Sheng, L., Chen, Z., Shao, J.: Thinking in Frequency: Face Forgery Detection by Mining Frequency-Aware Clues. In: Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII. p. 86–103. Springer-Verlag, Berlin, Heidelberg (2020)
26. Roy, R., Joshi, I., Das, A., Dantcheva, A.: 3D CNN Architectures and Attention Mechanisms for Deepfake Detection. In: Publishing, S.I. (ed.) Handbook of Digital Face Manipulation and Detection (2022)
27. Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M.: FaceForensics++: Learning to Detect Manipulated Facial Images. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 1–11 (2019)
28. Shiohara, K., Yamasaki, T.: Detecting Deepfakes With Self-Blended Images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18720–18729 (June 2022)
29. Tan, M., Le, Q.: EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In: Chaudhuri, K., Salakhutdinov, R. (eds.) Proceedings of the 36th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 97, pp. 6105–6114. PMLR (09–15 Jun 2019)
30. Verdoliva, L.: Media Forensics and DeepFakes: An Overview. IEEE Journal of Selected Topics in Signal Processing **14**(5), 910–932 (2020)
31. Wodajo, D., Atnafu, S.: Deepfake Video Detection Using Convolutional Vision Transformer. CoRR **abs/2102.11126** (2021)
32. Wu, X., Xie, Z., Gao, Y., Xiao, Y.: SSTNet: Detecting Manipulated Faces Through Spatial, Steganalysis and Temporal Features. In: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 2952–2956 (2020)
33. Yang, X., Li, Y., Qi, H., Lyu, S.: Exposing GAN-synthesized Faces Using Landmark Locations. In: Proceedings of the ACM Workshop on Information Hiding and Multimedia Security. p. 113–118. IH&MMSec'19, Association for Computing Machinery, New York, NY, USA (2019)
34. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. IEEE Signal Processing Letters **23**(10), 1499–1503 (2016)
35. Zhou, P., Han, X., Morariu, V.I., Davis, L.S.: Learning Rich Features for Image Manipulation Detection. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1053–1061 (2018)