Beyond Deepfakes: The Rise of Fully Synthetic Videos and Their Detection

Luisa Verdoliva

UNINA

24 June 2025

vera.ai final online webinar

Fully synthetic generated videos

Examples of videos generated from scratch giving a short description

Sora

“Historical footage of California during the gold rush”

Pika

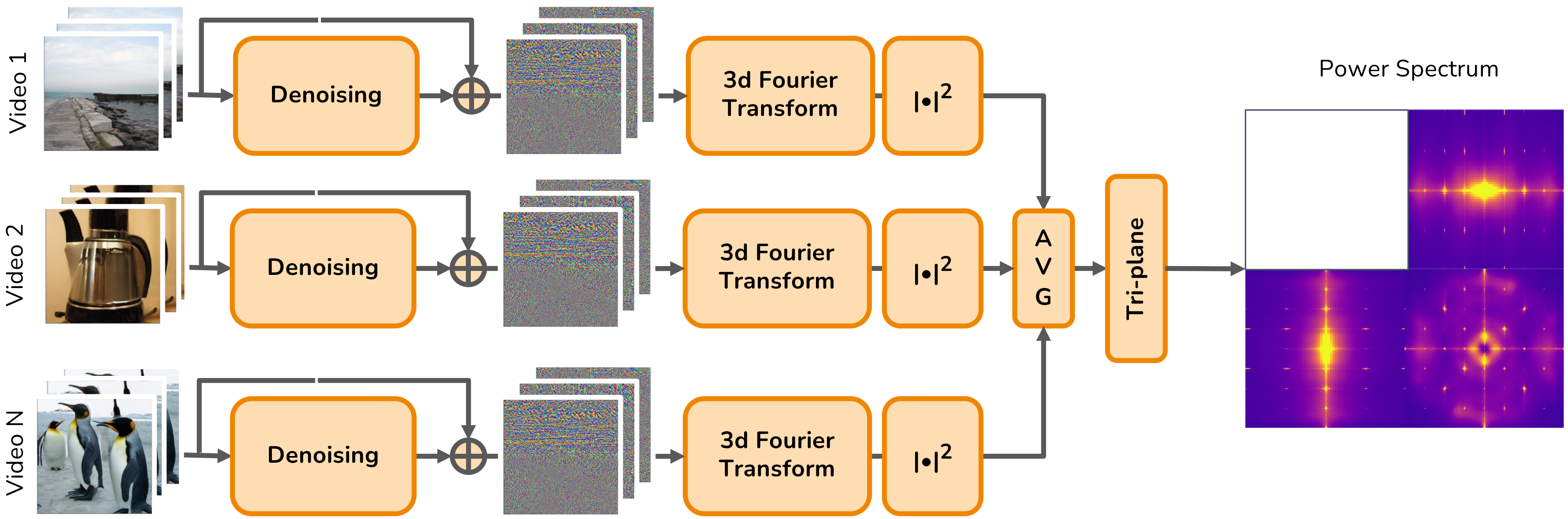
“cinematic shot, extreme close up dolly shot on a stylish japanese girl with dreads standing on a pink desert”

Runway ML

“an astronaut running through an alley in Rio de Janeiro”

3D Fourier analysis

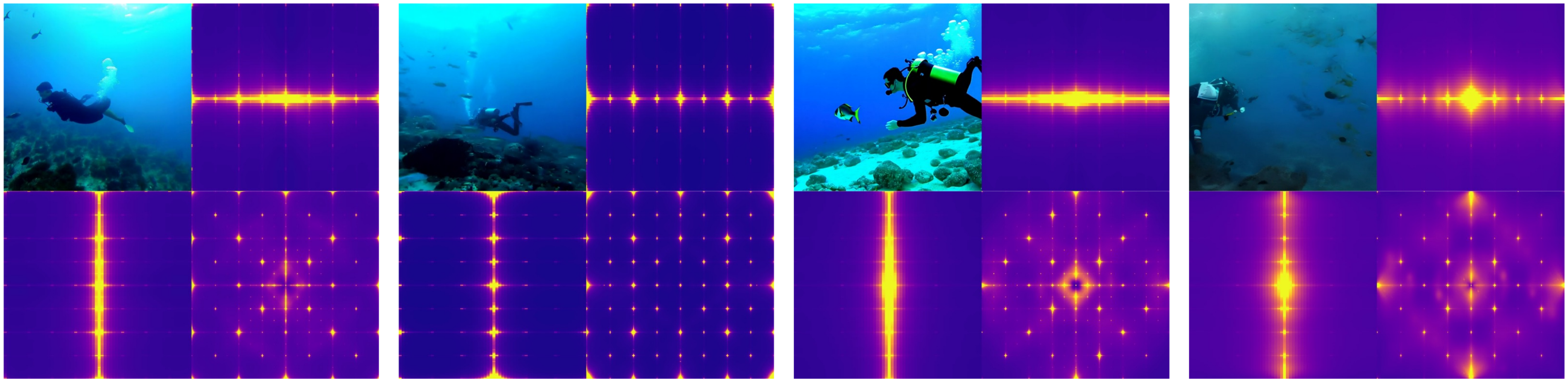
We analyzed the traces left from each generator by computing the power spectrum of the residual frames along different directions (xy, zy and xz)



Fingerprints in the Fourier domain

The spatial power spectra present typical spectral peaks (visible as bright spots) caused by the upsampling process in the generative architecture

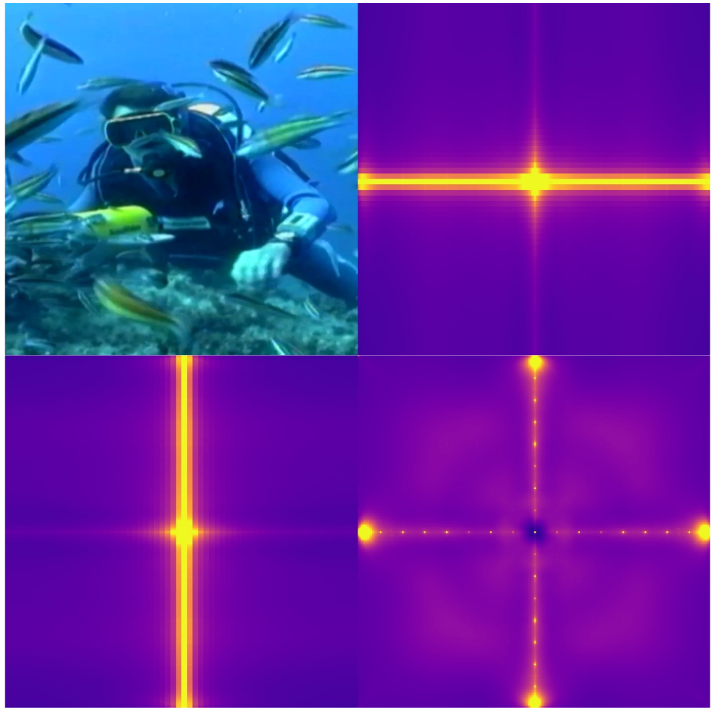
Similar peaks are also visible along the temporal direction



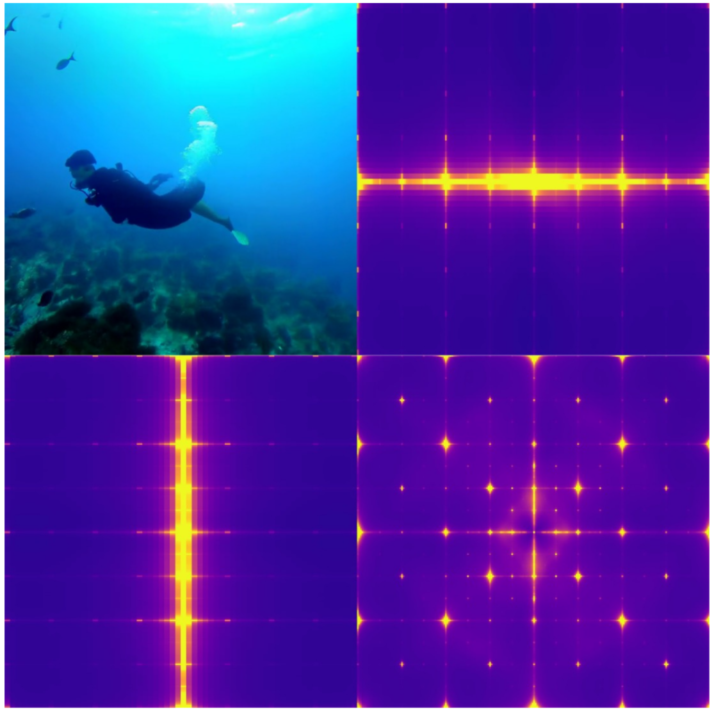
Pyramid Flow Mochi-1 Allegro Nova

AI-generated vs real videos

Such artifacts are not present in real videos which show compression-related traces



Real data

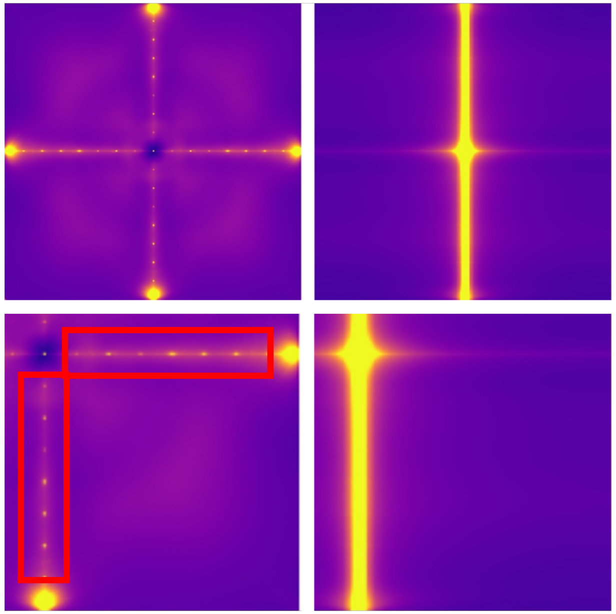


Synthetic data

Artifacts analysis after compression

Forensic clues are highlighted by circles, while peaks originated by compression are highlighted by red boxes

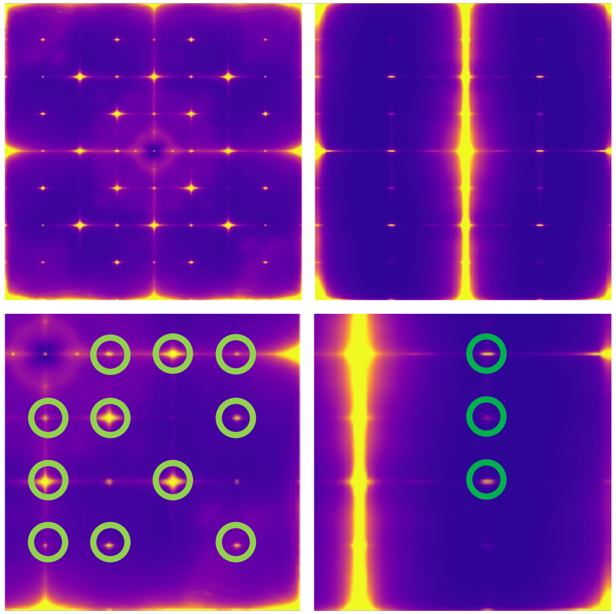
Compressed real video



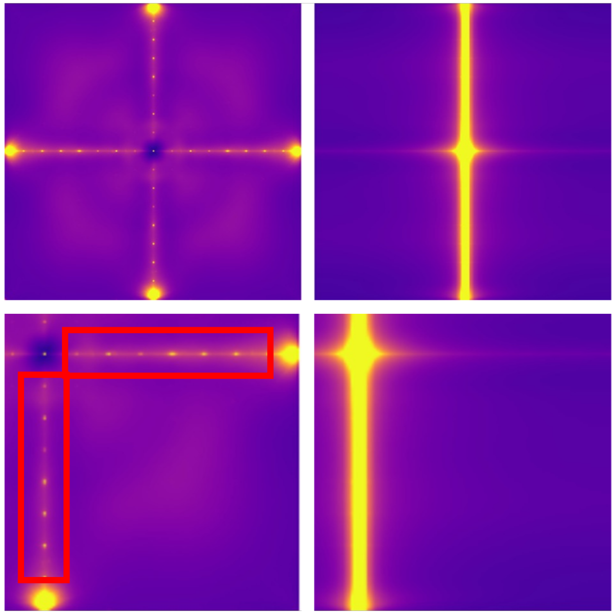
Artifacts analysis after compression

Forensic clues are highlighted by circles, while peaks originated by compression are highlighted by red boxes

Synthetic video before compression



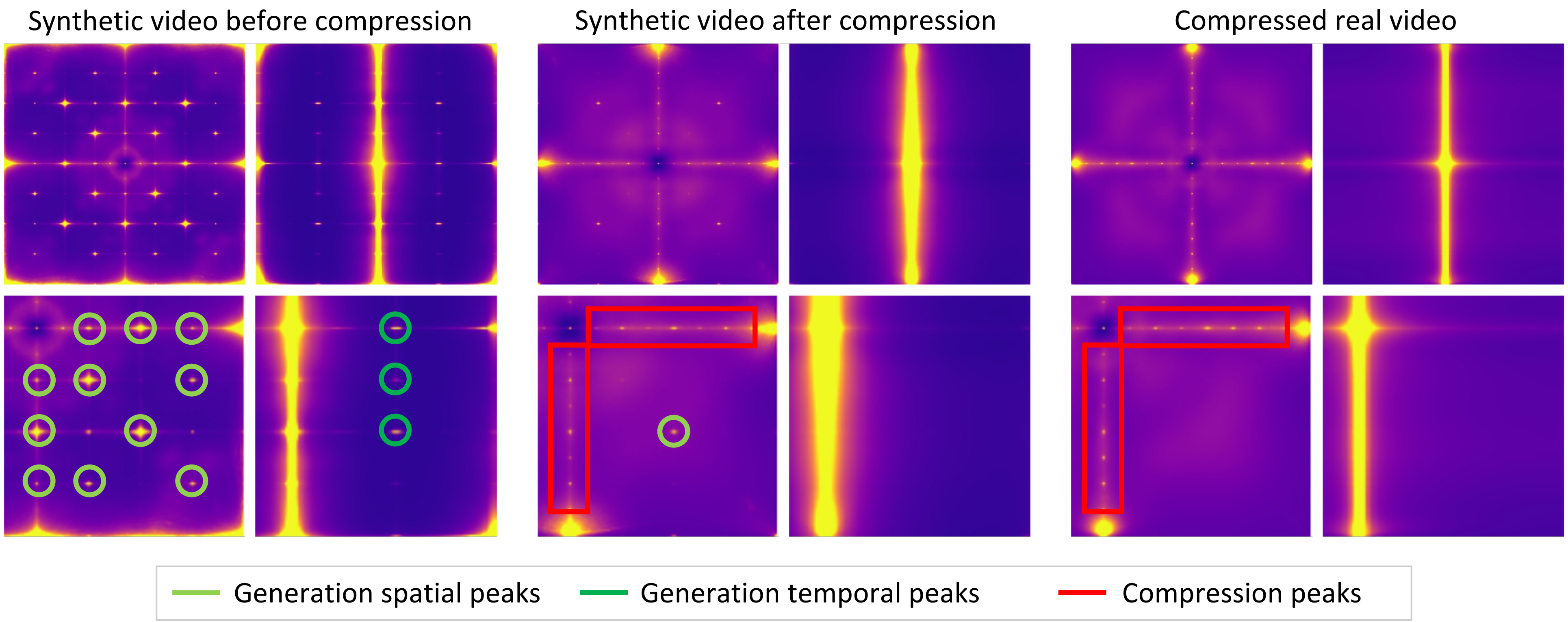
Compressed real video



Generation spatial peaks — Generation temporal peaks — Compression peaks

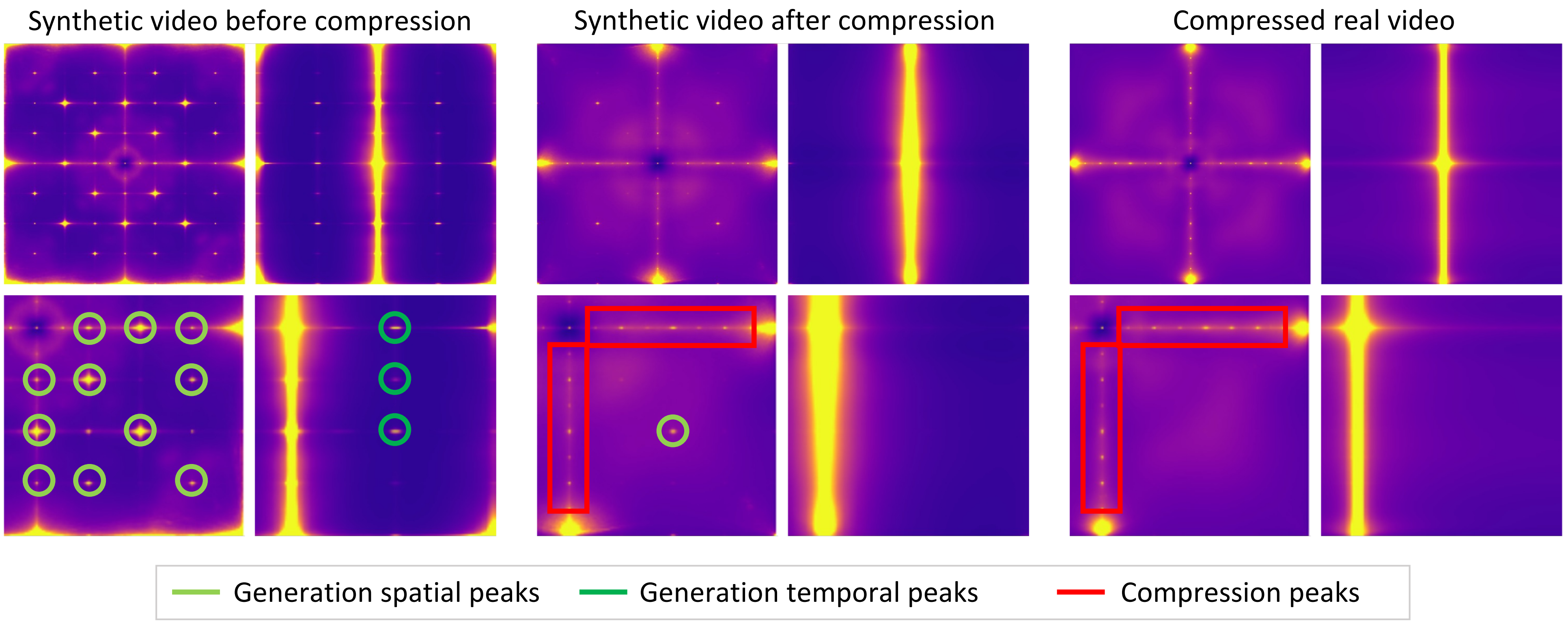
Artifacts analysis after compression

Forensic clues are highlighted by circles, while peaks originated by compression are highlighted by red boxes



Artifacts analysis after compression

Temporal forensic clues disappear after compression, while those along the diagonal spatial directions are still present



Considerations

Inconsistencies in the middle frequency content along the diagonal directions are more robust to commonly used video codecs

We propose an augmentation strategy that replaces specific frequency bands to guide the model to exploit more relevant forensic cues

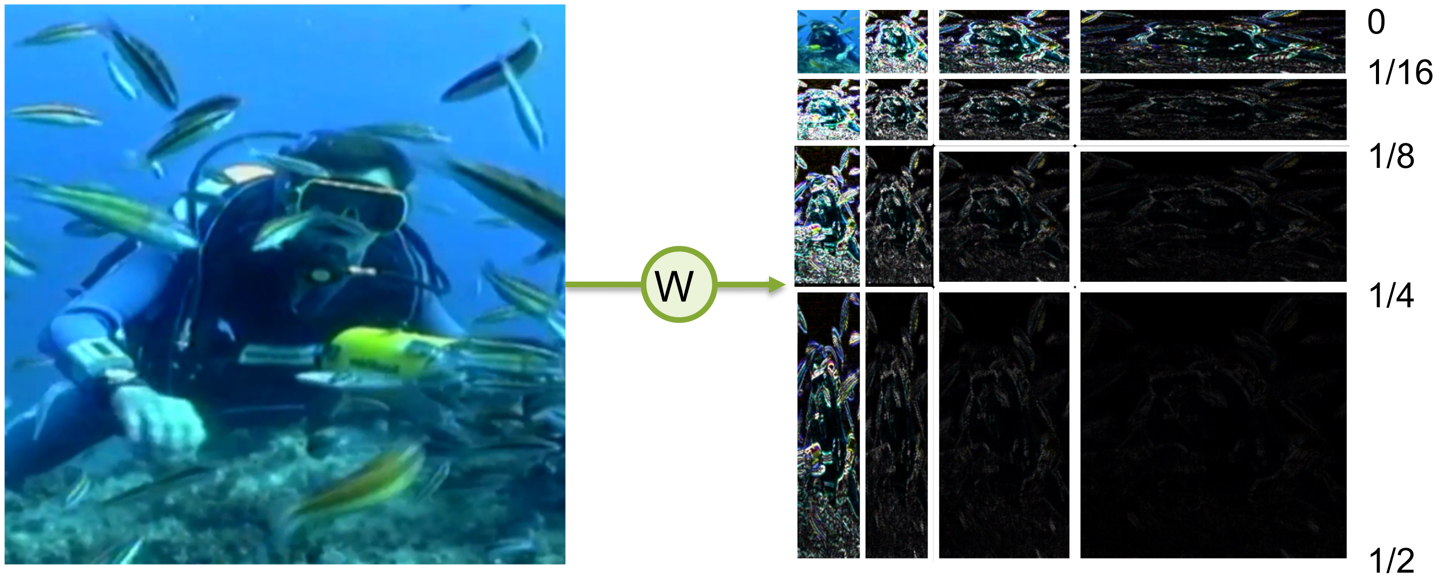
The proposed augmentation aims to avoid that the model polarizes on the horizontal/vertical frequencies

R. Corvi, D. Cozzolino, E. Prashnani, S. De Mello, K. Nagano, L. Verdoliva, “Seeing What Matters: Generalizable AI-generated Video Detection with Forensic-Oriented Augmentation” arXiv preprint arXiv:2506.16802, 2025

Wavelet transform

The Wavelet Transform decomposes the signal into several frequency-related sub-bands

We replace the low frequencies bands from real to fakes



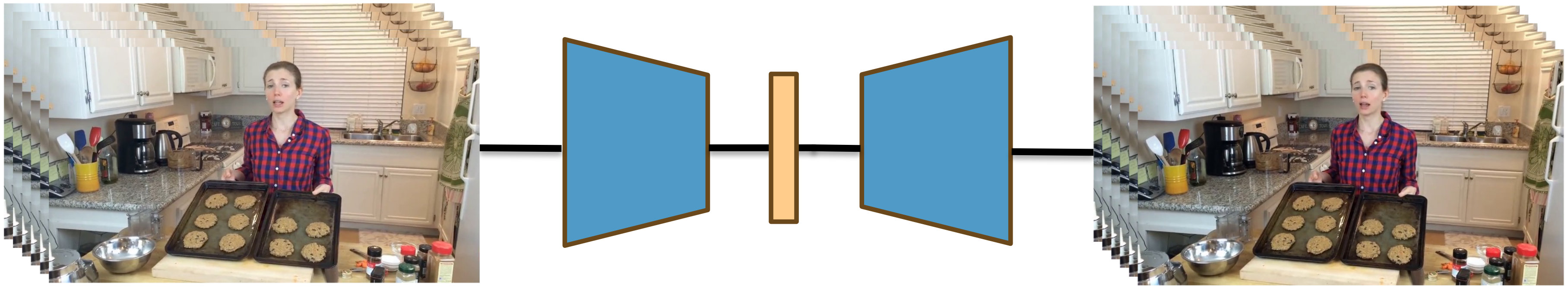
Spatial Domain

Wavelet Domain

Real vs Fake videos: alignment

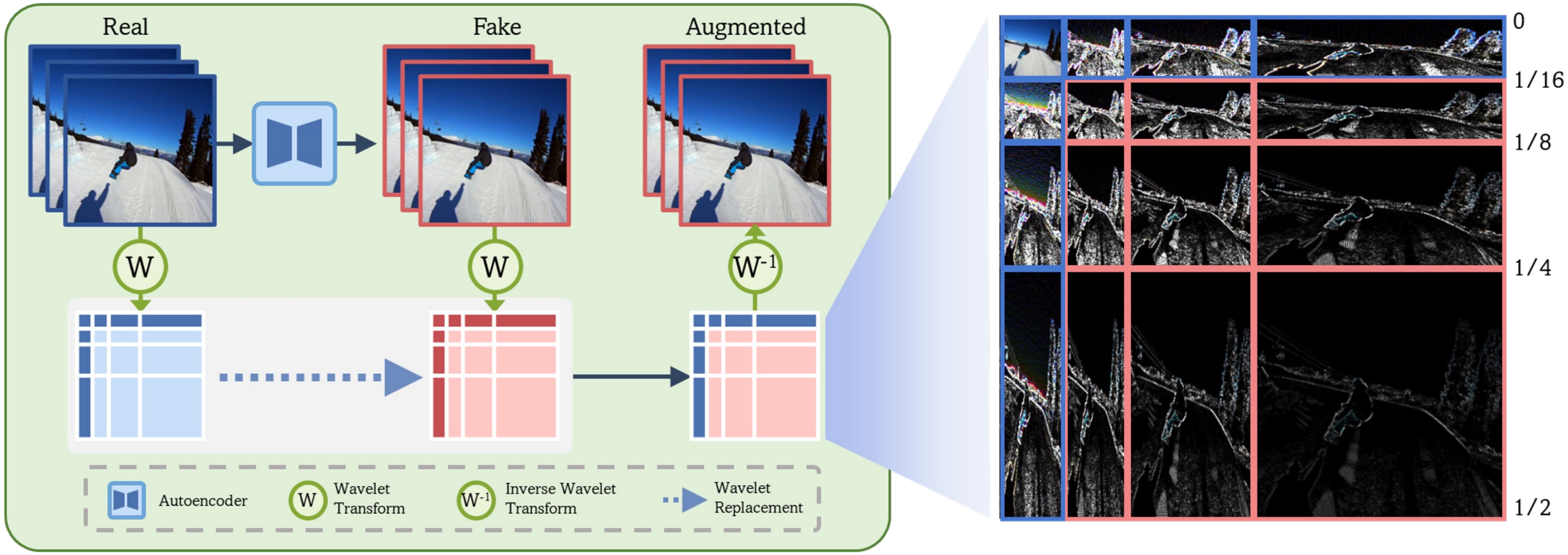
In order to replace the wavelet bands, we need an exact match of the semantic content between the real and synthetic videos

To align real and fake videos, we generate the synthetic content by passing the real videos through the autoencoder of a synthetic generator (i.e. Pyramid Flow)



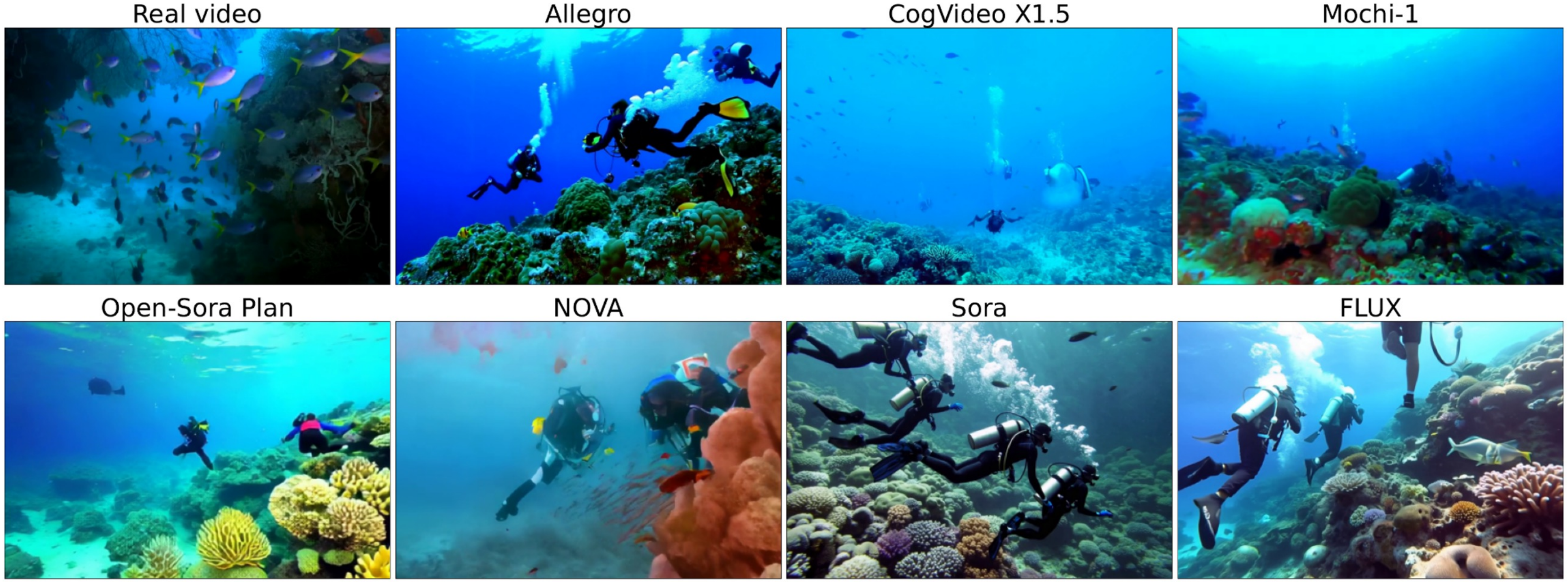
Augmentation strategy

We propose a wavelet-based augmentation strategy that encourages the model to learn frequency cues distinguishing real from synthetic content



Dataset of fully synthetic generated videos

We created a dataset of 10,000 AI-generated videos



"A group of scuba divers are swimming in a coral reef with colorful tropical fish."

Dataset of fully synthetic generated videos

vera.ai

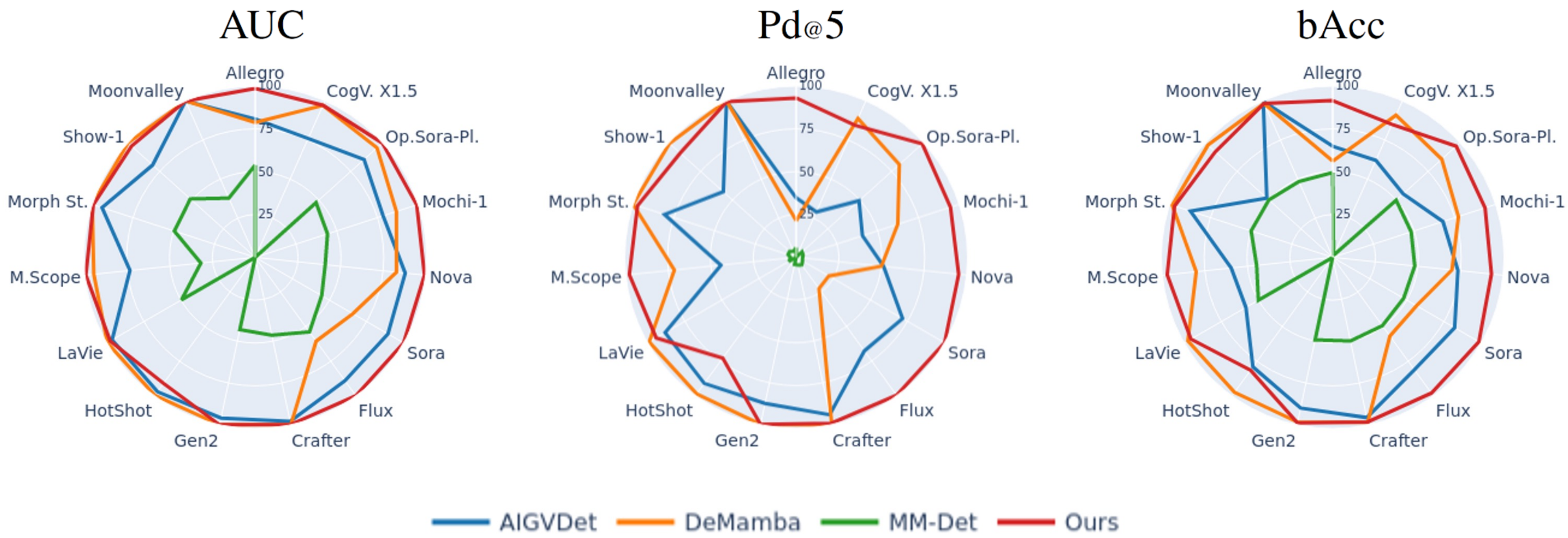
We used only state-of-the-art text-to-video generators that produce videos with high resolution, high frame rate and good results on VBench [1]

	Frame Rate	Resolution	Length	Quality Score (VBench)	Semantic Score (VBench)
Panda-70M (Real)	24 FPS	1280x720	5s-10s	-	-
Pyramid Flow	24 FPS	1280x768	5s	84.74%	69.62%
Allegro	15 FPS	1280x720	6s	83.12%	72.98%
Cogvideo X	15 FPS	1360x768	5s	82.78%	82.78%
Mochi-1	30 FPS	848x480	5s	82.64%	70.08%
Open-Sora Plan	18 FPS	640x532	5s	80.14%	65.62%

[1] Vbench leaderboard: https://huggingface.co/spaces/Vchitect/VBench_Leaderboard

Experimental results

Comparison with SoTA methods proposed for synthetic video detection on 16 generative models across different evaluation metrics



Conclusions

We propose a wavelet-based training augmentation that promotes learning more discriminative frequency cues to distinguish real from synthetic content

Our training paradigm improves the generalizability of the detector without the need for complex algorithms and large datasets that include multiple generators

Next step: develop a strategy that can exploit discriminative forensic cues present in the temporal domain

Luisa Verdoliva / UNINA

Contact: verdoliv@unina.it

Follow us on Twitter: @veraai_eu
Website: https://www.veraai.eu/

Co-financed by the European Union, Horizon Europe programme, Grant Agreement No 101070093.

Additional funding from Innovate UK grant No 10039055 and the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract No 22.00245