COMPROMIS

COntenus Multimédia PROtégés par Machines IntelligenteS

Protected Multimedia Contents using Intelligent Machines

Projet COMPROMIS
WP 2: Research activities in Integrity

Authors:
Romaric Audigier
Patrick Bas
Jan Butora
Emma Coletta
Christophe Charrier
Nicholas Evans
Emmanuel Giguet
Mohamed Mallat
Pierre-Alain Moellic
Benjamin Negrevergne
Anna Taylor
Massimiliano Todisco
Alexis Winter

Internal Reviewer:
Teddy Furon

April 23, 2026

Contents

1

Presentation of the whole project

The COMPROMIS project is based on a modern vision of multimedia data protection with deep learning at its core. This project defends the idea that the protection of multimedia data must necessarily be associated with the security of the tools that analyse this data, i.e. Artificial Intelligence nowadays. The observation is simple: Multimedia data protection is certainly the area of cybersecurity that has benefited the most from AI, but it has neglected the verification of the security level of this new tool. AI has become one of the weakest links in multimedia data protection.

Yet, the multimedia data protection community is the best suited to study the security of deep learning by mapping its cardinal values (integrity, confidentiality and identification) to deep learning data (training data, test data, learned model).

This parallelism between the protection of multimedia data and the security of deep learning creates a virtual cross-fertilisation that ensures end-to-end security, from the data to the algorithms that process it. The scientific challenges thus concern both the classic applications of multimedia data protection (forensic analysis, watermarking, biometrics, deepfakes, steganography, selective encryption) and the emerging field of deep learning. These scientific problems directly address the concerns of security agencies, industry, and multimedia content producers.

In the context described in the previous section, given the technological developments underway, the scientific challenges that this project intends to address are given by the cardinal values of multimedia data protection, namely:

1. Confidentiality: Confidentiality must be protected by encrypting multimedia data not only during storage and transmission but also during processing, including by Artificial Intelligence. Confidentiality is also understood to mean limiting the leakage of information concerning multimedia data. Indeed, when it concerns personal data (face, voice, etc.), confidentiality protects privacy in many cases (irreversibility of biometric signatures, for example).

2. Integrity: This value defends the authenticity of multimedia data, i.e. whether it has been intentionally manipulated by changing the semantic content or by inserting a hidden message. Authenticity analysis has to deal with the history of the data, which can be edited multiple times in its life in a benign way (compression, format transcoding, colorimetry change...). The challenge is therefore to discern intent (misappropriation or not?) and also to bring evidence of manipulation.

3. Identification / Ownership: Identification is understood in a broad sense: for example, knowing how to recognise a multimedia data (as a quasi-copy of an original version), or identifying its source or its rightful owner (the artist of a Work of art). In biometrics, identifying the source is equivalent to identifying the individual on the basis of his or her biometric traits.

2

Work package presentation

WP 2 – **Research activities in Integrity** – includes activities related to the verification of the integrity of multimedia data (forensic analysis) and associated models. The research is divided into two components involving research related to learning and to applications, and is organised into three distinct tasks. They are described in the following.

2.1 Learning

- Robustness to adversarial examples (T2.1 - resp. Miles)
  - The thesis '*Designing Adversarial Attacks for Randomized Neural Networks*' (resp. Miles) develops an attack whose effectiveness can be formally demonstrated for stochastic neural networks
  - The thesis '*Design of consistent surrogate loss functions for training robust neural networks*' (resp. Miles) designs new loss functions consistent with 0/1 loss in the adversarial setting
  - The Post-Doc and Engineering work '*Development of a continuous integration platform for testing and validating model robustness*' (resp. Miles) develops a testing platform quantitatively evaluating the robustness of models

2.2 Applications

- Deepfake (T2.2 - resp. EURECOM)
  - The thesis '*Deepfake generation and detection on residual errors*' (resp. GREYC) studies weak signals related to video time coding combined with digital forensic information
  - The thesis '*Detection and anti-detection of deepfakes using the sensor noise properties*' (resp. CRIStAL) differentiates deepfakes from natural data via the presence or absence of developed sensor noise, but also generates non-detectable data using the properties of the same noise
  - The thesis '*Deepfakes detection with audio-visual behavioral cues*' (resp. LinkMedia) exploits high-level information (behaviour, attitude, facial expressions, prosody...) characterising a known person to provide clues for the detection of deepfakes
  - The thesis '*Detection solutions for adversarial audio deepfakes*' (resp. EURECOM) investigates a new generation of adversarial attacks designed to fool both a voice biometric system and a deepfake/spoofing detector
- Integrity in Biometrics (T2.3 - resp. EURECOM)
  - The thesis '*Backdoor attacks against automatic speaker verification systems*' (resp. EURECOM) undertakes the first study of backdoor attacks on automatic speaker verification systems and also researches strategies for their detection
  - The thesis '*Integated solutions to deepfake/spoofing aware automatic speaker verification*' (resp. EURECOM) collects data from a larger number of speakers and allows not only further exploration of joint optimisation but also the development of individual classifiers that perform simultaneous detection and recognition
- Image analysis (T2.4 - resp. CEA)

– The thesis (resp. CEA) '*Defense of models for scene analysis against adversarial attacks*' proposes attacks and defences applicable to scene analysis modules, including those for object detection and object instance search in images

3

Executive summary

The COMPROMIS project addresses the integrity of multimedia data and the security of the artificial intelligence systems that analyse it, in a context where generative models, adversarial attacks, and large-scale learning paradigms increasingly threaten trust in digital content. Within WP2, the research activities have focused on advancing the scientific foundations, evaluation methodologies, and practical tools required to assess, detect, and mitigate integrity violations affecting images, audio, and learned models themselves. The main achievements reported in Section 4 span adversarial robustness, deepfake and forensic analysis, biometric integrity, and secure image analysis, and collectively contribute to a more reliable and explainable protection of multimedia contents in realistic threat settings.

A first set of contributions advances the understanding and evaluation of adversarial robustness in deep learning models. The project has identified fundamental limitations in existing robustness assessments, particularly for stochastic and randomized neural networks, where standard attack strategies lead to overly optimistic conclusions. New principled adversarial attacks have been developed to properly target such models, revealing vulnerabilities that were previously underestimated. In parallel, the project has explored black-box adversarial attacks, a more realistic threat model in which attackers interact with systems only through queries. By leveraging recent advances in derivative-free and manifold-based optimisation, the work demonstrates how attack efficiency can be significantly improved while reducing query complexity. To support rigorous and reproducible evaluation, a continuous integration platform for robustness testing has been developed, enabling systematic comparison of defences under standardised attack protocols. Together, these contributions strengthen the scientific basis for evaluating model robustness and provide practical tools to avoid biased or misleading security claims. In response to the rapid evolution of the field, the consortium is also preparing a new research direction linking adversarial examples with watermarking and generated content detection, with initial work initiated through a dedicated internship.

A second major line of work targets deepfake detection and multimedia forensic analysis, addressing both visual and speech modalities. On the image side, the project has produced state-of-the-art methods for localising generative inpainting and image manipulations, exploiting high-level semantic representations learned by large vision transformers. These approaches achieve fine-grained localisation of manipulated regions and remain robust to common post-processing operations such as resizing and compression. Complementary work has introduced noise-based deepfake detectors that explicitly preserve and exploit sensor and residual noise, achieving strong performance on high-quality generated images and providing interpretable cues about the origin of the decision. In addition, the project has advanced JPEG forensics through compatibility-based analysis, enabling precise localisation of manipulations at the block level and offering explainable evidence of tampering even after recompression. On the audio side, the project has conducted a systematic analysis of self-supervised speech representations for deepfake detection, revealing how artifacts introduced by text-to-speech and voice conversion attacks manifest differently across model layers. This work opens the way not only to robust detection, but also to source tracing and attack characterisation, a key requirement for forensic investigations.

The project has also made significant progress on biometric integrity, with a particular focus on voice biometrics. One contribution introduces Latent Secret Spin, a secure watermarking technique that embeds imperceptible, cryptographically bound information directly into the latent space of a neural audio codec. This method enables reliable authentication of speech integrity even after common signal processing operations, while remaining statistically indistinguishable from noise without the secret key. A second contribution addresses the need for explainable biometric decisions, proposing a speaker-dependent, phoneme-conditioned framework that detects spoofed or deepfake speech as deviations from learned bona fide speech distributions. By operating at the level of phonetic units and temporal regions, the approach provides intrinsic interpretability, identifying where and how generative artifacts affect biometric systems. Together, these works combine protection and explanation, reinforcing trust in biometric technologies under adversarial and generative threats.

Finally, the project has investigated image analysis security in modern learning paradigms, particularly in distributed and federated settings. One contribution provides the first in-depth analysis of backdoor attacks against federated adaptation of Vision Transformer models using parameter-efficient fine-tuning techniques such as LoRA. The work reveals how architectural constraints can significantly affect backdoor persistence and demonstrates that commonly used evaluation protocols may produce biased or misleading conclusions. This analysis contributes to more robust and fair assessment methodologies for federated learning security. Another contribution creates a benchmark to fairly compare the state-of-the-art adversarial attacks applied to object detection models. The study also investigates their transferability to different neural networks architectures, as well as their utilization to build the most effective adversarial training strategy for a robust defense. In parallel, the project has proposed a new loss-function framework for deep steganalysis explicitly optimised for very low false-positive rates, a critical requirement in real forensic applications. By integrating decision-theoretic principles directly into training, the proposed approach substantially outperforms standard losses under realistic operating conditions.

Overall, the results of WP2 significantly advance the state of the art in multimedia integrity and AI security. The project delivers new attack models, detection methods, evaluation tools, and explainable mechanisms that better reflect real-world threats faced by security agencies, industry, and content producers. By combining theoretical insights with practical, reproducible solutions across images, audio, biometrics, and learning systems, COMPROMIS contributes to strengthening trust in multimedia content and the intelligent machines that process it.

4

Main achievements

4.1 Robustness to adversarial examples (T2.1 - resp. Miles)

We present the contributions developed within the project to advance the understanding and evaluation of adversarial robustness in deep learning models. The work addresses fundamental limitations in existing robustness assessments, particularly in settings involving stochastic models, black-box access, and incomplete or biased evaluation protocols. Several of the research directions initially proposed in the project have progressed substantially, including the design of adversarial attacks for randomised neural networks and the development of a continuous integration platform for systematic robustness evaluation. In addition, rapid progress has been made towards more efficient black-box adversarial attacks based on derivative-free manifold optimisation. Together, these efforts provide more faithful threat models, reveal previously overlooked vulnerabilities, and contribute practical tools and perspectives for the rigorous and reproducible evaluation of model robustness.

4.1.1 Designing adversarial attacks for randomised neural networks

Randomised classifiers (i.e. classifiers that behave non-deterministically) have been proposed as a robust and theoretically grounded solution to design models that are robust to adversarial attacks. However, it remains unclear how to properly attack randomised models, and as a consequence, empirical validations of defence mechanisms are typically overestimated.

In earlier results published outside the context of the project, we have demonstrated that any randomised model can be modelled as an (possibly) infinite random mixture of classifiers. This has led us to focus our research on the specific problem of attacking randomised mixtures of classifiers, rather than attacking randomised classifiers in general. In this context we identified that existing methods designed for deterministic classifiers could not be attacked without losing *completeness*, i.e. the ability of an attack algorithm to discover an attack when there exist one. We also identify that other attack algorithms designed specifically for attacking mixtures suffer from limitations; in particular, they generate attacks that may be improved by targeting a superset of classifiers from the mixture.

Building on these observations, we designed a new attack, LCA, that demonstrates strong properties while remaining usable in practice. Empirical evaluation demonstrates that LCA can discover attacks that neither PGD-based strategies nor ARC can. These results has been published in [10].

4.1.2 Designing more efficient blackbox attacks using manyfold optimisation

Most research on adversarial examples has focused on white-box attacks, that is, attacks that assume full access to the target model, (including its weights and architecture). Although this is the most conservative setting, it is also unrealistic: many models are only accessible through online APIs, where interaction is limited to queries, and full access to the model is unavailable. To target such models, researchers have developed black-box attacks, typically based on derivative-free optimisation techniques. However, existing black-box attacks may be improvable, and thus the true level of threat these models face via black-box attacks remains poorly understood.

To advance research in this direction, Bastien Cavarretta (a Ph.D. student hired within the project) and his supervisors are investigating recent theoretical results in the field of derivative-free manifold optimisation. These techniques have the potential to drastically reduce the number of model queries required to generate an attack, increasing the potency of black-box attacks. In "*Complexity Guarantees and Polling Strategies for Riemannian Direct-Search Methods*" [4], Cavarretta et al. have analysed the performance of two alternative variants of a derivative-free optimisation algorithm. These variants arise from two polling strategies: *projected* and *intrinsic*.

This work found that the two approaches perform similarly as the manifold dimension varies, but that the intrinsic variant is significantly more efficient for manifolds with large codimensions.

In the near future, we will leverage insights from this theoretical analysis to devise new black-box attacks and corresponding defence mechanisms.

4.1.3 On the Vulnerability of Retrieval in High Intrinsic Dimensionality Neighborhood

This work investigates the vulnerability of the nearest-neighbor search, a pivotal tool in pattern analysis and data science, especially for multimedia content. The vulnerability is measured as the relative amount of perturbation an attacker needs to add to a dataset point to modify its proximity to a given query. The statistical distribution of the relative amount of perturbation is derived from simple assumptions, outlining the key factor that drives its typical values: The higher the intrinsic dimensionality, the more vulnerable the nearest-neighbors search becomes. Experiments on six large-scale datasets validate this model up to some outliers, which are explained as violations of the assumptions. See Figure 4.1. More information here [9].

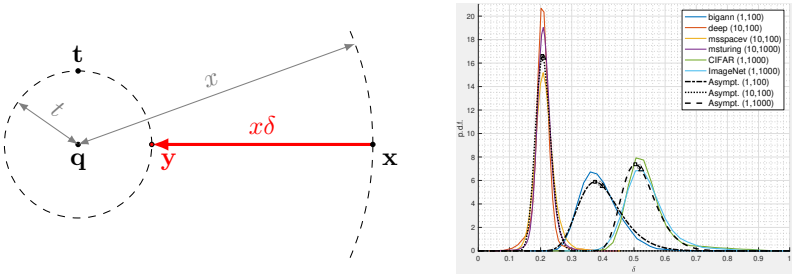


Figure 4.1: Although relatively simple, our mathematical model predicts the distribution of the distortion to delude the nearest neighbour search with accuracy on many databases.

4.1.4 Development of a continuous integration platform for testing and validating model robustness

The continuous integration platform has been developed and is used to test various defence mechanisms in a rigorous and systematic way, in the context of multiple pedagogical and research endeavours. Most recently, we used it to systematically evaluate defence mechanisms proposed by students of our master programme as well as other researcher of the teams. Participants registered on the platform and, by doing so, gained access to a Git repository that included stub code for implementing their defence mechanisms.

Every day, the platform automatically retrieves all repositories that have been updated since the previous day, performs a variety of attacks (based on AutoAttack), and reports the results via a website. Examples of these reports are shown in Figure 4.2.

4.2 Deepfake detection and forensic analysis (T2.2 - resp. EURECOM)

Our research addresses the growing threat posed by deepfake generation and advanced image and audio manipulations through the development of robust, explainable, and fine-grained forensic analysis methods. The contributions span both visual and speech modalities and target complementary levels of analysis, from local manipulation localisation to global authenticity assessment. On the image side, the work introduces state-of-the-art approaches for detecting generative inpainting, deepfake imagery, and JPEG-based manipulations by leveraging semantic or behavioral representations, noise residuals, and compatibility constraints inherent to imaging pipelines. On the audio side, we investigate how self-supervised speech representations encode deepfake artifacts across network layers, enabling not only accurate detection but also source tracing and attack characterisation. Together, these works advance deepfake forensics by combining principled modelling of generative artifacts with interpretable detection strategies that remain robust under realistic post-processing and distribution shifts.

4.2.1 Retinex-Guided Latent Refinement for Face Swapping

Face swapping is a computer vision technique that involves exchanging one person's face with another's in an image or video. While this technology has gained popularity through social media augmented reality filters, it also

This table was generated on 2025-12-13 at 06:21. See more results here. See last results here.

results										
project_name	group_name	hostname	status	time	time_per_image_ms	acc_nat	acc_pgdlinf	acc_pgdl2	agg	error_msg
jean-ponce	Master-IASD	upnquick	Success	5401.12	270.06	100.0	97.77	99.29	197.06	None
BestOf2023-1	profs	upnquick	Success	180.47	9.02	75.0	70.07	70.76	140.83	None
gradient-hackers	Master-IASD	upnquick	Success	793.77	39.69	67.5	70.03	68.64	138.66	None
blast_attack	Master-IASD	upnquick	Success	3009.92	150.5	61.25	58.16	69.06	127.21	None
attaqueoudefense	Master-IASD	upnquick	Success	4826.02	241.3	73.75	60.41	66.05	126.45	None
noeyedeer	Master-IASD	coktailjet	Success	3424.05	171.2	83.75	51.23	68.77	120.0	None
rattataque	Master-IASD	upnquick	Success	1174.61	58.73	88.75	48.59	71.16	119.75	None
exocet	Master-IASD	coktailjet	Success	1797.56	89.88	77.5	51.34	66.1	117.44	None
BestOf2024-1	profs	upnquick	Success	2782.26	139.11	68.75	51.56	63.72	115.28	None
BestOfMiles	profs	upnquick	Success	5413.08	270.65	76.25	52.08	62.84	114.92	None
the-taithon-canon	Master-IASD	upnquick	Success	1797.91	89.9	81.25	49.05	65.7	114.75	None
BestOf2024-2	profs	coktailjet	Success	1218.31	60.92	68.75	51.74	59.8	111.54	None
best_defense_is_attack	Master-IASD	coktailjet	Success	3218.2	160.91	75.0	50.92	59.24	110.16	None
counter_attack	Master-IASD	upnquick	Success	1701.71	85.09	63.75	46.1	63.14	109.24	None
nyc	Master-IASD	upnquick	Success	1371.58	68.58	70.0	44.39	63.95	108.33	None
attackonpixels	Master-IASD	coktailjet	Success	3801.11	190.06	81.25	44.34	52.51	96.85	None
neural-nightmare	Master-IASD	coktailjet	Success	2880.58	144.03	93.75	42.54	54.11	96.65	None
BestOf2023-2	profs	coktailjet	Success	113.09	5.65	62.5	40.84	53.52	94.36	None
invisible_attack	Master-IASD	coktailjet	Success	8265.97	413.3	100.0	31.59	51.88	83.47	None
the-advengers	Master-IASD	upnquick	Success	1286.85	64.34	68.75	31.61	50.56	82.17	None
jogabonito	Master-IASD	upnquick	Success	1372.33	68.62	68.75	31.37	48.86	80.23	None
attaquedestitans	Master-IASD	coktailjet	Success	145.42	7.27	87.5	18.67	38.39	57.06	None
attack_mesonet	Master-IASD	upnquick	Success	114.38	5.72	56.25	22.33	33.3	55.63	None
attackonnetworks	Master-IASD	upnquick	Success	114.03	5.7	62.5	18.3	32.57	50.87	None
base_model	profs	coktailjet	Success	113.1	5.65	56.25	6.02	25.12	31.14	None
attackus	Master-IASD	upnquick	Error	0	0	0	0	0	0	FileNotFoundError
attack-of-babrumen	Master-IASD	coktailjet	Error	0	0	0	0	0	0	AttributeError: '
ciclose-10	Master-IASD	upnquick	Error	0	0	0	0	0	0	FileNotFoundError
compo-4-3-3	Master-IASD	upnquick	Error	0	0	0	0	0	0	FileNotFoundError
harissa	Master-IASD	coktailjet	Error	0	0	0	0	0	0	FileNotFoundError
madraf	Master-IASD	coktailjet	Error	0	0	0	0	0	0	FileNotFoundError
team_joie	Master-IASD	upnquick	Error	0	0	0	0	0	0	RuntimeError: Err
troublemakers	Master-IASD	coktailjet	Error	0	0	0	0	0	0	FileNotFoundError

Figure 4.2: Example of automatic reports created by the continuous integration platform.

holds significant implications for cinema, video games, and digital security. Advances in both Generative Adversarial Networks (GANs) and diffusion models have enabled increasingly realistic face swaps, featuring precise details and

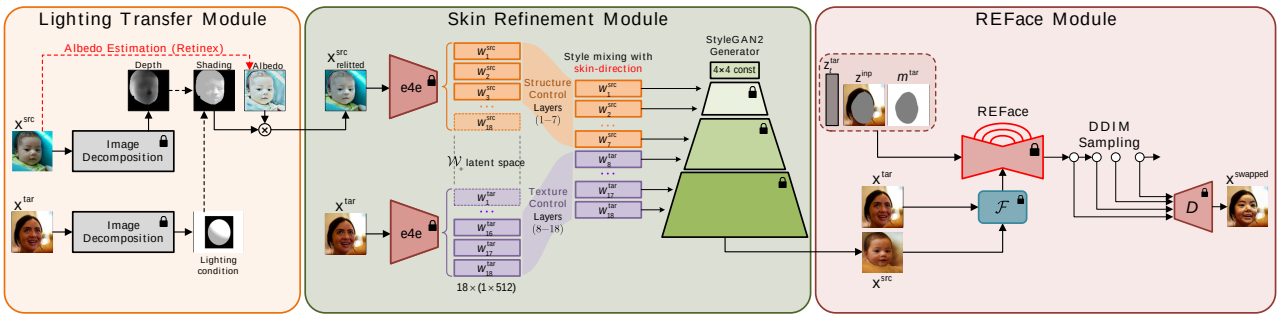


Figure 4.3: Overview of the three-stage Face Swapping architecture: featuring initial lighting alignment, StyleGAN2-based skin refinement, and final synthesis powered by the REFace model. First, the source and target images are processed through the Lighting Transfer Module to mitigate lighting inconsistencies. Next, the Skin Refinement Module performs latent space manipulation to align skin tones. Finally, these refined inputs are fed into the REFace Module, which uses a diffusion-based inpainting process to synthesize the high-fidelity final output, ensuring identity preservation and structural consistency.

refined facial expressions. However, results often suffer from lighting inconsistencies and mismatched skin tones between the source and target.

To address these issues, we developed a novel three-stage, training-free approach to face swapping designed to enhance the realism of images generated by diffusion models [13]. As depicted in Fig. 4.3, the first stage involves a lighting transfer module based on Retinex theory. This module decomposes the image to align the source image with the target illumination condition, mitigating lighting inconsistencies and ensuring a realistic blend before the main synthesis occurs.

The second stage introduces a skin refinement module that blends the source identity's structural layers with the target's texture and color layers in the disentangled latent space. This achieves accurate skin tone transfer while preserving the source's unique identity. To address the lack of explicit skin tone labels, we propose a novel method that leverages 3D Morphable Model (3DMM) coefficients and K-means clustering to define a precise manipulation direction within the latent space.

Finally, we used the REFace approach as a frozen diffusion-based synthesis engine. It leverages pretrained weights to generate the final high-fidelity swapped result without requiring additional training or fine-tuning (see Figure 4.4.

Experimental results on the CelebAMask-HQ dataset demonstrate a significant improvement in visual quality, achieving a reduced FID score of 7.16. Although identity fidelity is slightly lower than the baseline model (0.498 vs. 0.632 in ID similarity), the system offers a superior overall trade-off by outperforming existing methods in photometric realism and the preservation of target attributes such as pose and expression.

4.2.2 Generative inpainting localisation

We introduce DinoLizer, a DINOv2-based model for localizing manipulated regions in generative inpainting. Our method builds on a DINOv2 model pretrained to detect synthetic images on the B-Free dataset. We add a linear classification head on top of the Vision Transformer's patch embeddings to predict manipulations at a 14×14 patch resolution. The head is trained to focus on semantically altered regions, treating non-semantic edits as part of the original content. Because the ViT accepts only fixed-size inputs, we use a sliding-window strategy to aggregate predictions over larger images; the resulting heatmaps are post-processed to refine the estimated binary manipulation masks. Empirical results show that DinoLizer surpasses state-of-the-art local manipulation detectors on a range of inpainting datasets derived from different generative models. It remains robust to common post-processing operations such as resizing, noise addition, and JPEG (double) compression. On average, DinoLizer achieves a 12% higher Intersection-over-Union (IoU) than the next best model, with even greater gains after post-processing. Our experiments with off-the-shelf DINOv2 demonstrate the strong representational power of Vision Transformers for this task. Finally, extensive ablation studies comparing DINOv2 and its successor, DINOv3, in deepfake localisation confirm DinoLizer's superiority.

See Figure 4.5. More information here [7].

4.2.3 JPEG forensics using compatibility tests

Given a JPEG pipeline (compression or decompression), this paper shows how to find the antecedent of a 8×8 block. If it exists, the block is *compatible* with the pipeline. For unaltered images, all blocks are always compatible

Figure 4.4: Comparison of the visual superiority of the proposed Face Swapping method (labeled *Ours*) against the REFace baseline (labeled *Baseline*). The baseline often results in an unnatural look due to lighting and skin tone mismatches. In contrast, the proposed pipeline produces seamless results with strong photometric consistency and accurate skin tone matching, particularly in challenging cases with extreme lighting or diverse skin colors.

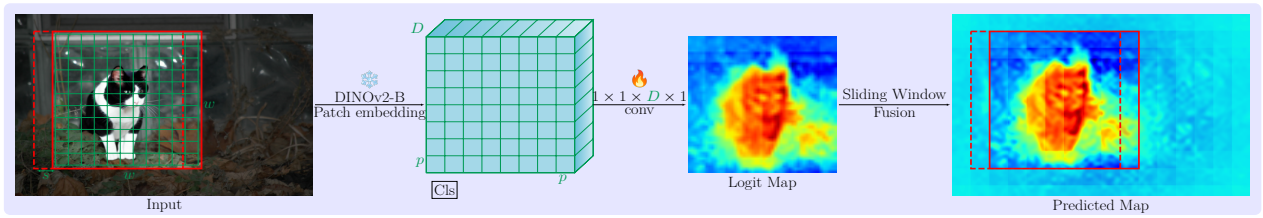


Figure 4.5: Generative inpainting localisation - Principle of DinoLizer: the image is decomposed into a set of 504×504 overlapping crops which are fed to the DINOv2 model to provide embeddings of dimension d for each patch plus a class token which is not considered here. A 1×1 trainable convolutional layer is used to infer a logit map, which is then fused with other overlapping maps in order to provide, after thresholding, a localisation mask.

with the original pipeline; however, for manipulated images, this is not always the case. This article demonstrates the potential of compatibility concepts for JPEG image forensics. It addresses the main challenge of finding a block antecedent in high-dimensional space. This solution relies on a local search algorithm with restrictions on the search space. We show that inpainting, copy-move, or splicing applied after a JPEG compression can be turned into three different mismatch problems and be detected. In particular, when the image is re-compressed after the modification, we can detect the manipulation if the quality factor of the second compression is higher than the first one. Our method can pinpoint forgeries down to the JPEG block, with high detection power and no false positives. We compare our method with two state-of-the-art models on localizing inpainted forgeries after a simple or a double compression. We show that under our working assumptions, it outperforms those models for most experiments.

See Figure 4.6. More information here [14].

(a) Original JPEG image at QF80 (b) Forged JPEG image at QF85 (c) Detection on decompressed JPEG

Figure 4.6: JPEG forensics using compatibility tests - Demonstration of finding incompatible JPEG blocks. The chimney was removed using in-painting (*i.e.* uncompressed data), and birds were spliced with different Quality Factors (QFs), one of them is fully aligned with the JPEG grid but not the other. The window was copy-moved aligned on the JPEG grid but not fully aligned, *i.e.* portions of blocks belong to the original image.

4.2.4 DeepFake Detection based on Noise Residuals

Deepfakes pose major challenges, especially regarding fraud, misinformation, and evidence tampering. As deepfakes become more prevalent, effective detection methods are essential. We introduce DJIN, a deepfake detection model that retains noise components by avoiding pooling layers in the initial stages. Pre-trained on ImageNet for steganography using the JIN version, DJIN outperforms CoDE and CLIP and is the best among all detectors mentioned for the In-Distribution dataset. DJIN is highly effective in handling high-quality images and processing images of various sizes. Since deepfake generators typically produce high-quality outputs, an explainability analysis reveals that DJIN leverages image noise by focusing on darker areas in real images and brighter areas in generated ones. The code is publicly available at https://gitlab.cristal.univ-lille.fr/mtdoi/djin.

See Figure 4.7. More information here [6].

4.2.5 Layer-wise embedding analysis for speech deepfake detection and source tracing

Speech models derived using self-supervised learning (SSL) are powerful tools for audio deepfake detection. Previous work has shown that lower SSL layers encode mainly stylistic information (e.g. voice identity), while higher layers better represent linguistic content. Attacks based on text-to-speech (TTS) synthesis and voice conversion (VC) produce different cues that distinguish deepfakes from bona fide speech. We therefore hypothesise that although

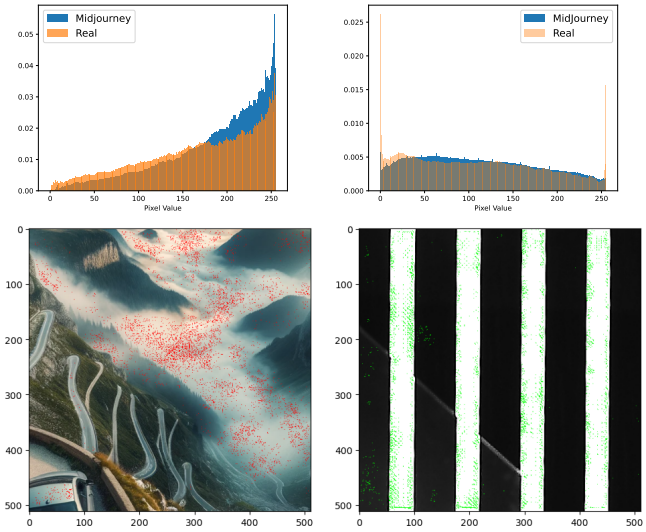


Figure 4.7: DeepFake Detection based on Noise Residuals - Probability of the top 1% most influential pixels according to Integrated Gradients w.r.t the total pixels of each value (0-255) (top left). Histogram of pixels' values in real images and Midjourney (top right). Integrated Gradients of midjourney sample image (bottom left), real image (bottom right).

embeddings from higher layers usually yield the best average detection performance, those from lower layers may remain informative for specific attacks. Moreover, combining representations from multiple layers may provide an attack *signature*, enabling not only deepfake detection but also identification of the attack type or underlying generative technology. To investigate how detection-relevant information evolves across an SSL architecture, we conducted a systematic analysis of embeddings extracted from different transformer layers of a WavLM model. Classification heads trained on individual layers show that TTS and VC attacks exhibit distinct artifacts and layer-dependent profiles. Canonical correlation analysis (CCA) experiments on the ASVspoof 5 dataset [18] reveal decreasing correlation with raw acoustic features from lower to higher layers, and show that middle layers are more effective for detecting TTS attacks, while VC attacks are better detected using higher-layer representations. A multi-layer approach to detection and source tracing is currently being prepared for submission to Odyssey 2026.

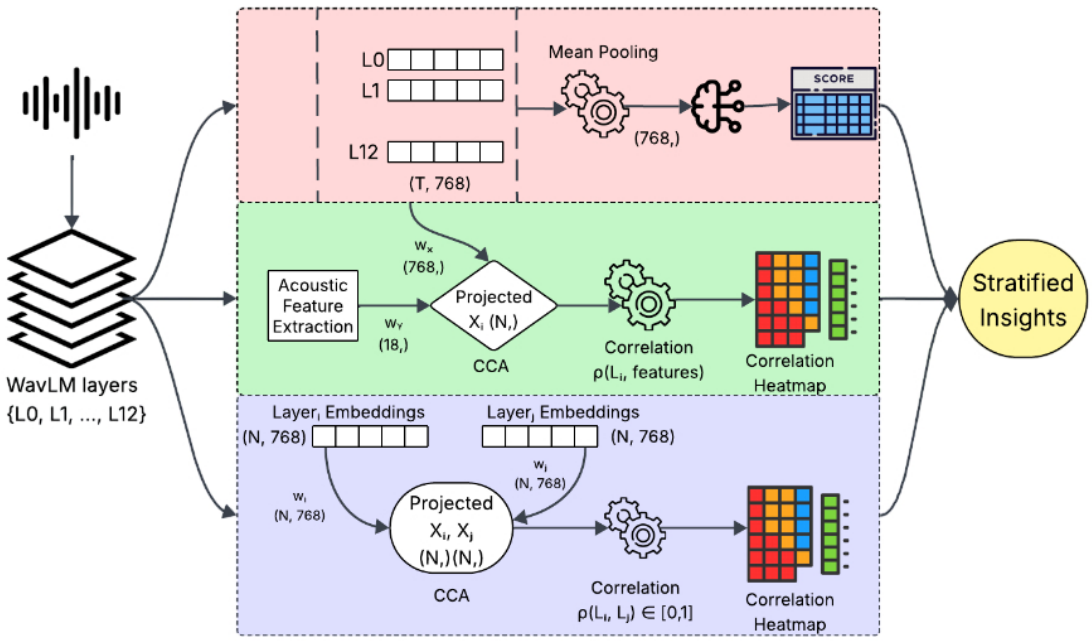


Figure 4.8: Layer-wise embedding analysis for speech deepfake detection.

4.2.6 Are Facial Action Units discriminative features to detect deepfakes?

The democratization of deepfakes necessitates detectors that remain effective against unseen attacks and robust to media degradation. While many detectors rely on low-level noise artifacts, these "weak signals" are often fragile. This work investigates whether Facial Action Units (FAUs) constitute discriminative features for deepfake detection. The rationale is twofold: generative models often struggle to consistently reproduce complex physiological muscle dynamics, and unlike pixel-level noise, FAUs are high-level semantic features potentially more resistant to compression or resizing. The limitation is that it can only be applied to well-known celebrities, for whom many true videos are available to build a model of their personal FAU. To determine whether FAUs are truly distinguishable features, we conduct a rigorous analysis across different learning configurations, comparing One-Class (trained only on real data) and Binary SVMs. We evaluate their utility for discerning real videos from deepfakes across multiple scenarios using the FakeAVCeleb dataset, aiming to determine whether FAUs offer a generalized and robust physiological signature for detection. This work is in progress.

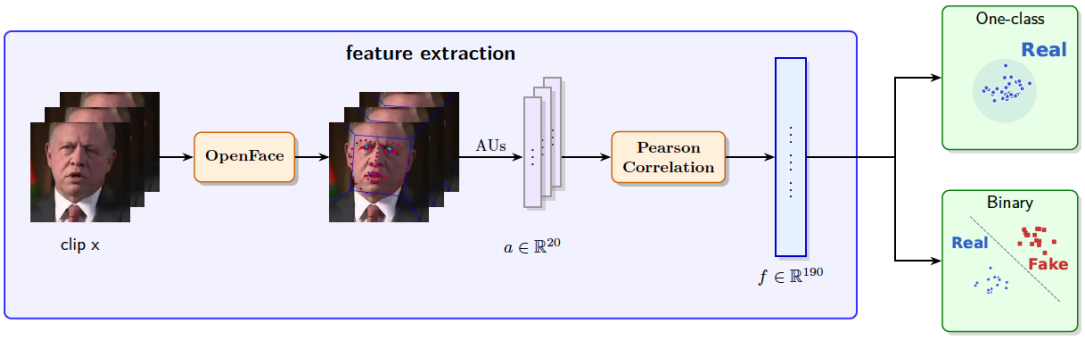


Figure 4.9: Facial Action Units pipeline for deepfake detection.

4.3 Integrity in biometrics (T2.3 - resp. EURECOM)

We address the integrity of biometrics by developing complementary methods that both protect and explain biometric decision-making in the presence of manipulation and generative attacks. The first contribution introduces Latent Secret Spin (LSS), a secure watermarking strategy designed to authenticate the integrity of speech signals. LSS embeds imperceptible, cryptographically bound watermarks directly into the latent space of a pretrained neural audio codec, enabling reliable detection even after common signal processing operations, while remaining statistically indistinguishable from noise without the secret key. This approach provides strong guarantees against unauthorised detection and forgery, making it well-suited to protect the integrity of biometric recognition. The second contribution focuses on explainable voice biometrics, proposing a speaker-dependent, phoneme-conditioned framework that detects deviations from human speech at the phonetic level. By leveraging discrete, phoneme-aligned representations learned through self-supervised speech models, the method identifies spoofed or deepfake speech as outliers relative to a learned bona fide distribution, while simultaneously highlighting the phonetic units and temporal regions responsible for the decision. Together, these two lines of work advance biometric integrity by combining secure signal authentication with intrinsically interpretable detection, offering both protection against manipulation and insight into how and where generative speech technologies affect biometric systems.

4.3.1 Latent secret spin

We introduce Latent Secret Spin (LSS), a secure watermarking strategy to provide for the authentication of speech integrity using imperceptible signals embedded directly into the latent feature space of a pretrained Encodec [8] model. By projecting these features into a principle component analysis (PCA) space, our method identifies specific planes to apply tiny rotations that encode information without degrading audio quality. A defining characteristic is a strict cryptographic binding: a secret key generates the unique rotation schedule, rendering the watermark close to impossible to replicate without the correct secret key. To any observer lacking the key, the embedded signal remains statistically indistinguishable from random noise. Watermark detection is performed by identifying the unique, key-dependent fingerprint introduced during the embedding stage. Results derived from a subset of bona fide samples extracted from the ASVspoof 5 [18] dataset are highly promising, demonstrating robust detection accuracy even

after audio manipulations such as MP3 compression and frequency filtering. Furthermore, detection performed with incorrect keys yields results equivalent to random guessing, confirming resilience against unauthorised detection or forgery. Future work will focus on scaling the evaluation, expanding data variability, and stress-testing the system against stronger attacks. Our work, which started in September 2025, is under preparation for submission to Odyssey 2026.

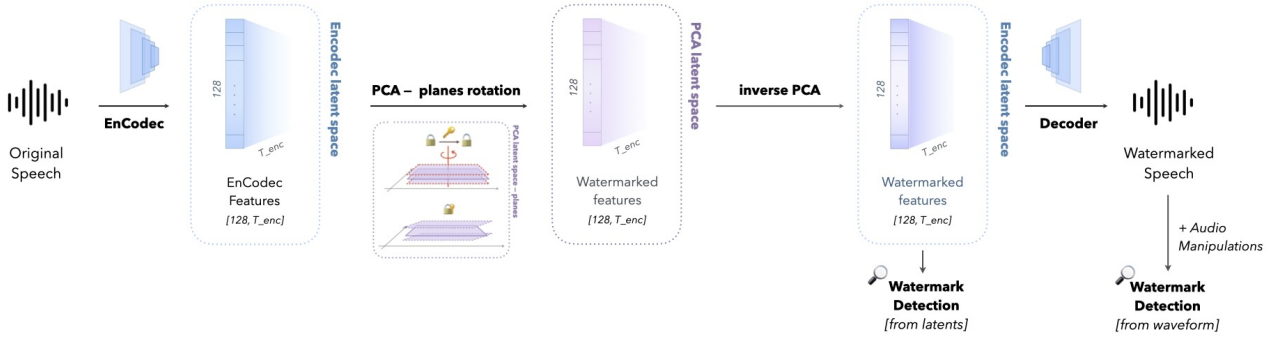


Figure 4.10: Latent secret spin (LSS) is used to embed imperceptible watermarks into recordings of speech. The key is speaker dependent, hence the approach is suited to the protection of biometric (voice) integrity.

4.3.2 Explainable voice biometrics through phoneme-conditioned plausibility

We introduce an approach to explainable voice biometrics which identifies deviations from human speech through speaker-dependent, phoneme-level analysis. Prior work has shown that discrete latent speech representations learned without phonetic supervision nevertheless exhibit strong correspondence with human-annotated phonemic categories, allowing them to serve as symbolic proxies for acoustic–phonetic states [1, 11]. Building on this insight, we employ a self-supervised speech encoder that produces phoneme-aligned discrete unit emissions [11], enabling segmentation and aggregation of speech into unit-level representations without reliance on text, pronunciation lexicons, or forced alignment. Our method is trained in a one-class setting using only bona fide speech from each speaker. For a given speaker, the model learns a compact bona fide region (speaker-specific distribution) in an attention-pooled embedding space derived from the unit emissions. At inference, a test utterance is scored by its deviation from the learned bona fide region: bona fide samples are expected to lie inside/near this region, whereas spoofed/deepfake samples tend to appear as outliers. An attention mechanism applied over the discrete units identifies the phonetic categories and time regions that contribute most to the score, yielding an intrinsically interpretable detection process. Rather than relying on global spectral artifacts, the proposed approach evaluates phoneme-conditioned acoustic plausibility and temporal consistency, enabling localisation of synthesis abnormalities at the level of individual phonetic units while maintaining biometric integrity through speaker-dependent modelling. This methodology supports robust detection of speech synthesis and voice conversion, and provides insight into which phonetic components and temporal dynamics of a particular speaker are most affected by specific generative models. Our work, which started in November 2025, is under preparation for submission to Odyssey 2026.

4.4 Image analysis (T2.4 - resp. CEA)

We study image analysis systems by investigating vulnerabilities and robustness issues arising in modern learning paradigms used for visual understanding and forensic security. The work focuses on three complementary directions. The first investigates the existing adversarial attacks applied to the object detection models and tries to answer the following three key questions: (1) How can we create a fair benchmark to impartially compare attacks? (2) How well do modern attacks transfer across different architectures, especially from Convolutional Neural Networks to Vision Transformers? (3) What is the most effective adversarial training strategy for robust defense? The second examines the resilience of large Vision Transformer models. The first work examines their robustness against adversarial example attacks when serving as foundation models. This means that they are public backbones used for diverse downstream tasks. In other contexts, the ViT models are tuned through federated learning. The second work focuses on backdoor attacks that can be enabled during distributed, parameter-efficient fine-tuning. By analysing how architectural constraints such as low-rank adaptation influence the persistence and evaluation of backdoors, this line of work highlights critical biases in current assessment protocols and contributes to more reliable security evaluations. The third direction targets deep steganalysis under realistic forensic conditions, proposing a

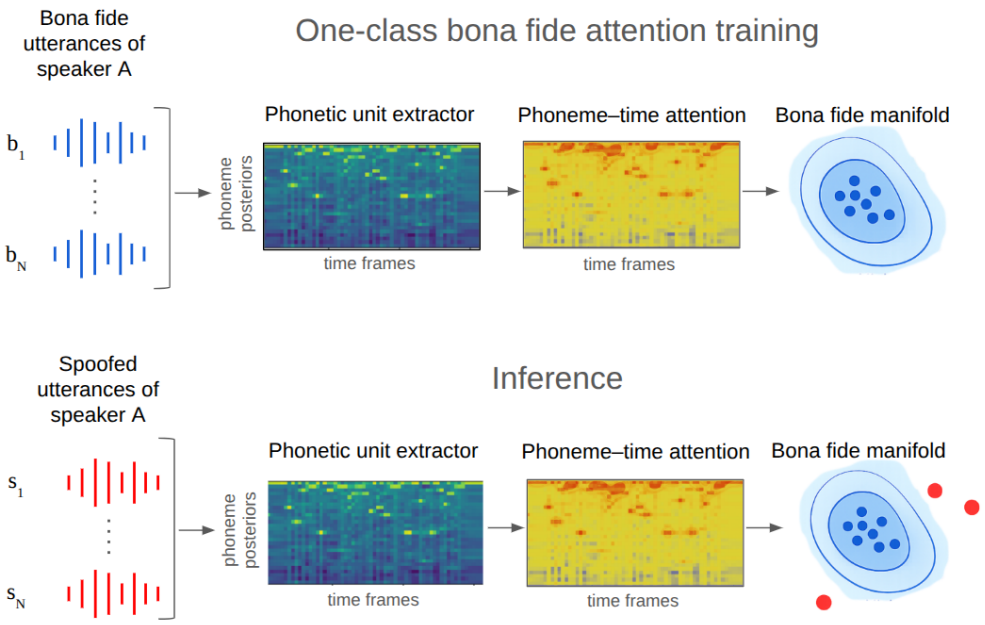


Figure 4.11: Phoneme-conditioned plausibility uses attention pooling on phoneme-aligned discrete unit emissions from a self-supervised speech encoder to score test utterances by their deviations from the speaker's bona fide distribution with a phoneme-time interpretation.

principled loss-function design that explicitly optimises performance at very low false-positive rates. Together, these contributions strengthen image integrity analysis by combining a better understanding of adversarial threats in distributed vision models with improved detection reliability for hidden information in digital images.

4.4.1 Defense of models for scene analysis against adversarial attacks

Object detection models are critical components of automated systems, such as autonomous vehicles and perception-based robots, but their sensitivity to adversarial attacks poses a serious security risk. Progress in defending these models lags behind classification, hindered by a lack of standardized evaluation.

Attacks and models can be classified based on their characteristics, context, and intent, as shown in Figure 4.12. The outcomes of attacks, i.e., the types of impact that the adversarial perturbation has on the predictions of the target detector, are then used to classify the attacks in our survey. An example of an inference result for each outcome is shown in Figure 4.13.

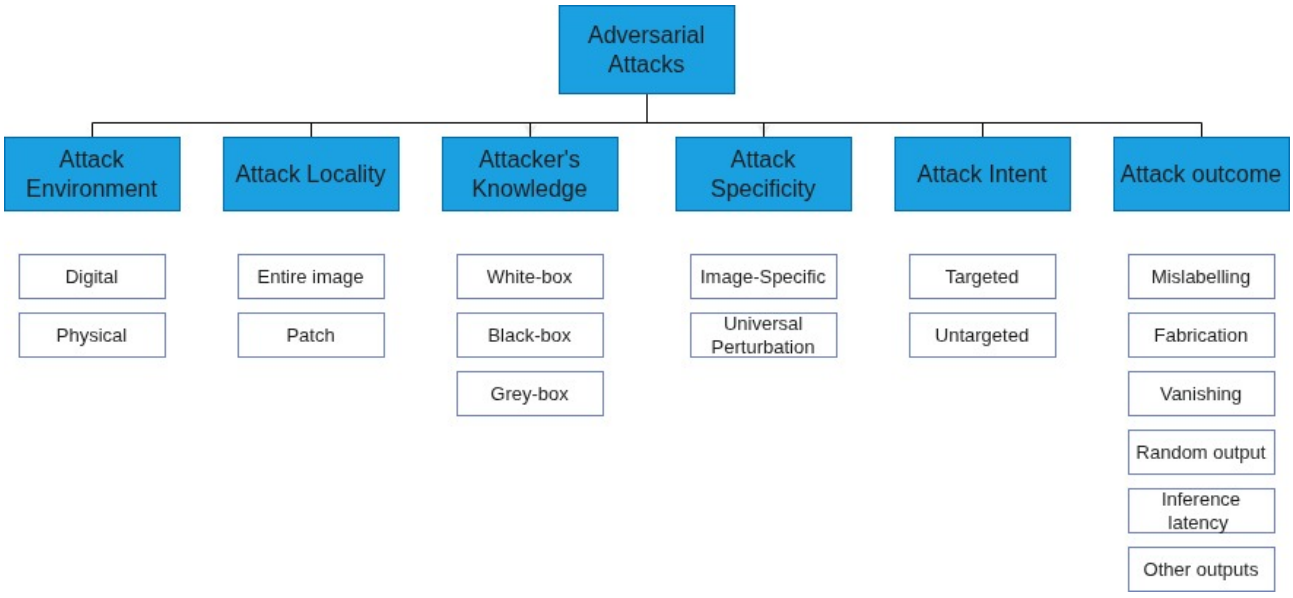


Figure 4.12: A taxonomy of adversarial attacks in object detection.

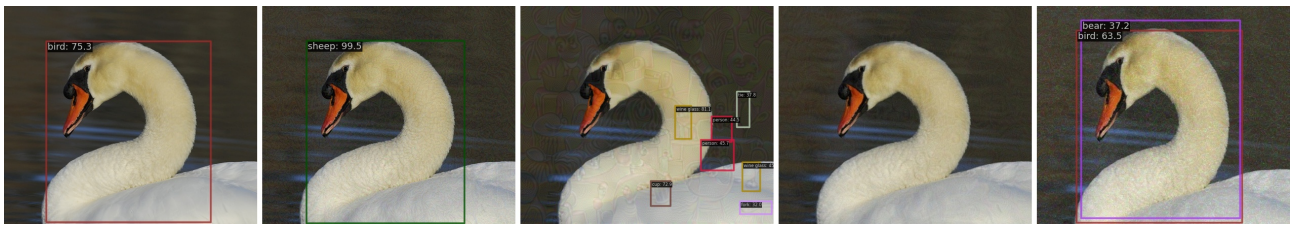


Figure 4.13: Inference results for different attack outcomes. From left to right: clean image, object mislabeling attack (EBAD [3]), random output attack (OSFD [5]), object vanishing attack, and object fabrication attack (PhantomSponges [16]).

However, it is nearly impossible to thoroughly compare attack or defense methods, as existing work uses different datasets, inconsistent efficiency metrics, and varied measures of perturbation cost. Indeed, choosing different levels of perturbation strongly impact on the attack effectiveness, but also on its perceptibility, as shown in Figure 4.14, where greater perturbation levels generate visible patterns on the image.

The present work addresses this gap by investigating three key questions: (1) How can we create a fair benchmark to impartially compare attacks? (2) How well do modern attacks transfer across different architectures, especially from Convolutional Neural Networks to Vision Transformers? (3) What is the most effective adversarial training strategy for robust defense? To answer these, we first propose a unified benchmark framework focused on digital, non-patch-based attacks. This framework introduces specific metrics to disentangle localization and classification errors and evaluates attack cost using multiple perceptual metrics.

Using this benchmark, we conduct extensive experiments on state-of-the-art attacks and a wide range of detectors. Our current findings reveal two major conclusions: first, modern adversarial attacks against object detection models show a significant lack of transferability to transformer-based architectures. Second, we demonstrate that the most robust adversarial training strategy leverages a dataset composed of a mix of high-perturbation attacks with different objectives (e.g., spatial and semantic), which outperforms training on any single attack.

More details and figures are available on the manuscript submitted to a journal [19].

4.4.2 Evaluating Backdoor Attacks Against Federated ViT Model Adaptation

Large models adaptation through Federated Learning (FL) addresses a wide range of use cases and is enabled by Parameter-Efficient Fine-Tuning (PEFT) techniques such as Low-Rank Adaptation (LoRA) [12]. LoRA relies on the fact that, after the fine-tuning, the difference between the initial and the new parameters is a low-rank sparse matrix. Formally, if $W_0 \in \mathbb{R}^{m \times n}$ is a pre-trained weight matrix, LoRA surrogates the updates for a low-rank decomposition $\Delta W_0 = AB$, where $A \in \mathbb{R}^{m \times r}$ and $B \in \mathbb{R}^{r \times n}$, $r << min(m, n)$ being the rank. The main advantage is that only A and B need to be updated at training time. However, this distributed learning paradigm faces several security threats, particularly to its integrity, such as backdoor attacks that aim to inject malicious behaviour during the local training steps of certain clients [20]. We propose the first analysis of the influence of LoRA on state-of-the-art backdoor attacks targeting Visual Transformer (ViT) model adaptation in FL. Specifically, we focus on backdoor lifespan (i.e., how long does the backdoor persist after the poison injection phase?), a critical characteristic in FL, that can vary depending on the attack scenario and the attacker's ability to effectively inject the backdoor. A key finding in our experiments is that for an optimally injected backdoor, the backdoor persistence after the attack is longer when the LoRA's rank is lower. With our analysis, we highlighted two key findings, illustrated in Fig. 4.15 (ViT pretrained on ImageNet and adapted on EuroSat with a standard FL setting: 100 participants, including 5 adversarial clients, from which 10 clients are randomly selected at each round):

1. By significantly reducing the model's capacity, LoRA slows down the learning of the backdoor. If the attacker is time-constrained in effectively injecting it, a lower r will make the backdoor less persistent. However, evaluating its lifespan under such conditions is misleading, as it is heavily biased by a short duration of the injection phase.

2. Under optimal backdoor injection conditions, applying LoRA with lower ranks actually increases the backdoor's lifespan (as opposed to the previous behaviour under partial injection), models indeed overwrite it more slowly if the rank is lower.

Importantly, this work highlights evaluation issues of backdoor attacks against FL (i.e., biased or misleading evaluations in SotA works such as [20]) and contributes to the development of more robust and fair evaluations of backdoor attacks in such a federated adaptation context. This work has been presented at the 18th International Symposium on Foundations & Practice of Security (FPS) [17], the code is publicly available (`https://gitlab.emse.fr/securityml/lora_backdoor_fl`).

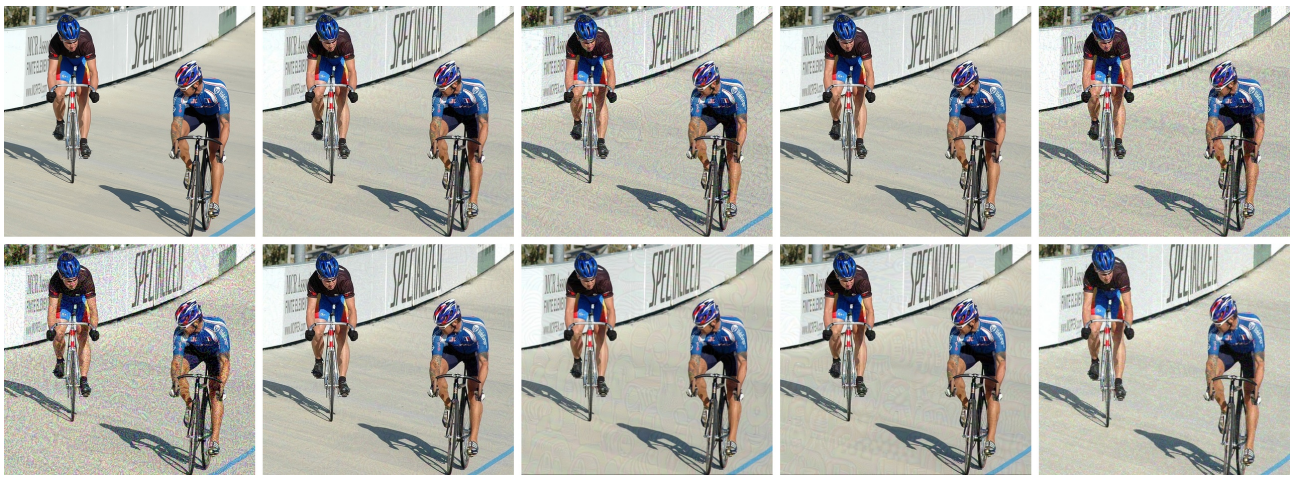
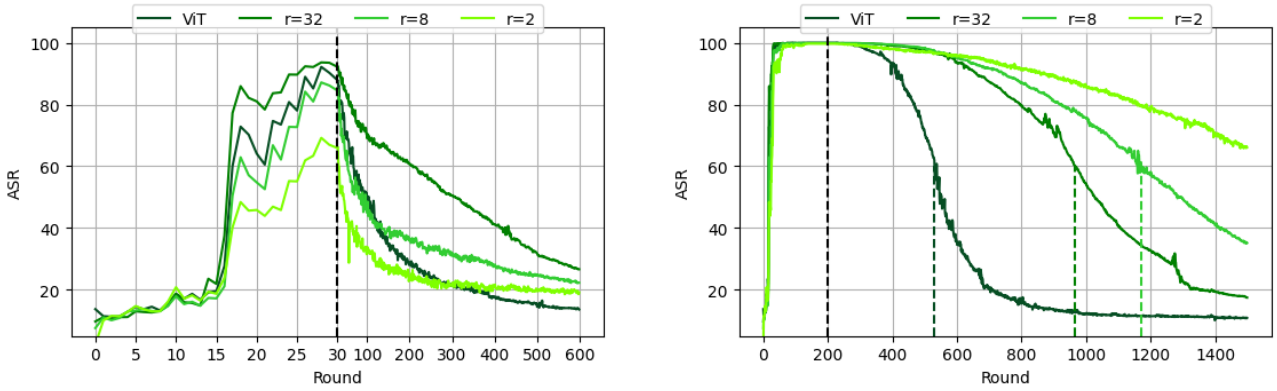


Figure 4.14: Visual samples from VOC2007 test set and their perturbed versions. Top row (left to right): benign image, CAA ($\epsilon = 10, 30$), and EBAD on YOLOv3 ($\epsilon = 10, 30$). Bottom row (left to right): EBAD on YOLOv3 ($\epsilon = 50$), EBAD on Faster R-CNN ($\epsilon = 10$), OSFD on YOLOv3 and Faster R-CNN ($\epsilon = 5$), and PhantomSponges ($\epsilon = 70$). Perturbations best viewed with zoom at 300%.



(a) Attack window = [0,30]

(b) Attack window = [0,200]

Figure 4.15: (Left) Attack Success Rate (ASR) for the ViT with and without LoRA. Note that we adopt a non-linear x-axis scale to improve the readability of the injection phase during the attack window. Injection is not optimal and, at the very end of the attack window, the lower the rank, the lower the ASR, which impacts the backdoor lifespan. (Right) With an optimal attack window, the behaviour is different: a lower rank induces higher persistence.

4.4.3 Task-Agnostic Attacks Against Vision Foundation Models

The study of security in machine learning mainly focuses on downstream task-specific attacks, where the adversarial example is obtained by optimizing a loss function specific to the downstream task. At the same time, it has become standard practice for machine learning practitioners to adopt publicly available pre-trained vision foundation models, effectively sharing a common backbone architecture across a multitude of applications such as classification, segmentation, depth estimation, retrieval, question answering and more. The study of attacks on such foundation models and their impact on multiple downstream tasks remains vastly unexplored. This work proposes a general framework that forges task-agnostic adversarial examples by maximally disrupting the feature representation obtained with foundation models. We extensively evaluate the security of the feature representations obtained by popular vision foundation models by measuring the impact of this attack on multiple downstream tasks and its transferability between models. See Figure 4.16. More information here [15].

4.4.4 Loss Function for Deep Steganalysis at Low False Positive Rate

Deep steganalysis has been crucial for detecting hidden messages in digital media for nearly a decade. However, its common security evaluation criterion-the probability of error under equal prior-fails to reflect real forensic challenges. In practice, low False Positive (FP) rates matter most, but are only adjusted empirically post-training. Standard

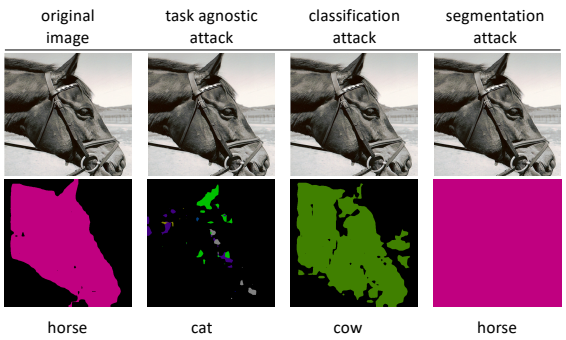


Figure 4.16: The task-agnostic attack deludes both the classification and the segmentation models, where the targeted attacks only corrupt the output of their targeted model.

classifiers, trained with cross-entropy loss, optimise balanced error rates rather than minimizing FPs. We propose a framework that integrates the likelihood ratio test into the loss function to optimise deep classifiers for low FP rates. Our method outperforms standard cross-entropy and other modern approaches, as demonstrated on the BOSSBase dataset across FP rates of 10^{-3} to 10^{-1} in both uncompressed and JPEG domains. Code available here https://github.com/janbutora/MODE-loss

See Figure 4.17. More information here [2].

Loss	α_0	P_{E}	$P_{\mathrm{D}}(\alpha)$		
			10^{-3}	10^{-2}	10^{-1}
CE	-	**0.1520**	0.2250	0.5440	**0.7830**
DTP [13]	0	0.1710	0.3595	0.5600	0.7445
Pat&Mat [1]	10^{-2}	0.1593	0.3090	0.5690	0.7720
	10^{-3}	0.1607	0.3180	0.5590	0.7730
MODE	10^{-2}	0.1578	0.3295	0.5710	0.7815
	10^{-3}	0.1618	**0.3695**	**0.5830**	0.7715

Figure 4.17: Loss Function for Deep Steganalysis at Low False Positive Rate - Average probability of error (P_E) and true positive rates (P_D) for various false positive rates (α) and loss functions: cross-entropy (CE), Deep Top Push (DTP) Pat&Mat and the proposed MODE loss. UERD steganographic embedding at JPEG QF100.

Bibliography

[1] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: a framework for self-supervised learning of speech representations. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc.

[2] Jan Butora and Patrick Bas. MODE: Loss Function for Deep Steganalysis at Low False Positive Rate. In *The 33rd European Signal Processing Conference (EUSIPCO)*, Palermo, Italy, September 2025.

[3] Zikui Cai, Yaoteng Tan, and M. Salman Asif. Ensemble-based blackbox attacks on dense prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[4] Bastien Cavarretta, Florentin Goyens, Clément W. Royer, and Florian Yger. Complexity guarantees and polling strategies for riemannian direct-search methods, 2025.

[5] Xinlong Ding, Jiansheng Chen, Hongwei Yu, Yu Shang, Yining Qin, and Huimin Ma. Transferable adversarial attacks for object detection using object-aware significant feature distortion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

[6] Minh Thong Doi, Jan Butora, Vincent Itier, Jérémie Boulanger, and Patrick Bas. DeepFake Detection based on Noise Residuals. In *GRETSI'25*, Strasbourg (67000), France, August 2025.

[7] Minh Thong Doi, Jan Butora, Vincent Itier, Jérémie Boulanger, and Patrick Bas. DinoLizer: Learning from the Best for Generative Inpainting Localization. working paper or preprint, November 2025.

[8] Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression, 2022.

[9] Teddy Furon. On the Vulnerability of Retrieval in High Intrinsic Dimensionality Neighborhood. *IEEE Transactions on Information Forensics and Security*, 20:3576–3586, 2025.

[10] Lucas Gnecco, Benjamin Negrevergne, and Yann Chevaleyre. Lattice Climber Attack: Adversarial attacks for randomized mixtures of classifiers. In *Proceedings of ECML PKDD 2025*, Porto, Portugal, September 2025. extended version with supplementary material.

[11] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021.

[12] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[13] Thu Hien Le, Christophe Charrier, Emmanuel Giguet, and Maxime Bérubé. Retinex-guided relighting and latent-space refinement for realistic diffusion-based face swapping. In *Media Watermarking, Security, and Forensics Conference, at IS&T Electronic Imaging 2026*, Burlingame, CA, USA, March 2026.

[14] Etienne Levecque, Jan Butora, and Patrick Bas. Dual JPEG Compatibility: a Reliable and Explainable Tool for Image Forensics. working paper or preprint, April 2025.

[15] Brian Pufler, Yury Belousov, Vitaliy Kinakh, Teddy Furon, and Slava Voloshynovskiy. Task-Agnostic Attacks Against Vision Foundation Models. In *5th Workshop of Adversarial Machine Learning at CVPR 2025*, pages 1–18, Nashville, United States, June 2025.

[16] Avishag Shapira, Alon Zolfi, Luca Demetrio, Battista Biggio, and Asaf Shabtai. Phantom Sponges: Exploiting non-maximum suppression to attack deep object detectors. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2023.

[17] Bastien Vuillod, Pierre-Alain Moellic, and Jean-Max Dutertre. Watch out for the lifespan: Evaluating backdoor attacks against federated model adaptation. *International Symposium on Foundations and Practice of Security*, 2025.

[18] Xin Wang, Hector Delgado, Hemlata Tak, Jee weon Jung, Hye jin Shim, Massimiliano Todisco, Ivan Kukanov, Xuechen Liu, Md Sahidullah, Tomi Kinnunen, Nicholas Evans, Kong Aik Lee, and Junichi Yamagishi. ASVspoof 5: crowdsourced speech data, deepfakes, and adversarial attacks at scale. In *The Automatic Speaker Verification Spoofing Countermeasures Workshop (ASVspoof 2024)*, pages 1–8, 2024.

[19] Alexis Winter, Jean-Vincent Martini, Romaric Audigier, Angélique Loesch, and Bertrand Luvison. Benchmarking adversarial robustness and adversarial training strategies for object detection. *Submitted to IJCV*, 2026.

[20] Hangfan Zhang, Jinyuan Jia, Jinghui Chen, Lu Lin, and Dinghao Wu. A3fl: Adversarially adaptive backdoor attacks to federated learning. *Advances in neural information processing systems*, 36:61213–61233, 2023.