

UNIVERSITÉ DE LILLE
FACULTÉ DES SCIENCES ET TECHNOLOGIES

Contre-mesures et méthodes de détection des fausses images et vidéos générées par des GAN

Stevenson PATHER

Master Informatique

Master mention Informatique



DÉPARTEMENT D'INFORMATIQUE
Faculté des Sciences et Technologies

juillet, 2025

Ce mémoire satisfait partiellement les pré-requis du module de Mémoire de Master, pour la 2^e année du Master mention Informatique.

Candidat: Stevenson PATHER, N° 11807967,
stevenson.pather.etu@univ-lille.fr

Encadrant: Marc TOMMASI, marc.tommasi@univ-lille.fr



DÉPARTEMENT D'INFORMATIQUE
Faculté des Sciences et Technologies
Campus Cité Scientifique, Bât. M3 extension, 59655 Villeneuve-d'Ascq

juillet, 2025

Résumé

Ces dernières années, les progrès rapides des modèles génératifs ont bouleversé notre rapport à l'image numérique. Les réseaux antagonistes génératifs (GANs) permettent aujourd'hui de produire des images et des vidéos d'un réalisme tel qu'elles deviennent difficilement distinguables du réel. Bien que ces techniques ouvrent de nouvelles perspectives dans divers domaines, elles soulèvent aussi des enjeux majeurs en matière de désinformation, de falsification visuelle et de confiance dans les contenus numériques.

Ce mémoire s'intéresse aux différentes méthodes de détection d'images et de vidéos générées par GANs, mais commence par interroger la perception humaine et les implications sociales de la prolifération de ces contenus. Ce cadrage initial permet de justifier l'importance des outils de vérification et des contre-mesures techniques.

Nous présentons ensuite un rappel des manipulations visuelles antérieures à l'ère des GAN, avant d'introduire les principes fondamentaux des modèles génératifs modernes. La majeure partie de ce travail est consacrée à l'étude des approches de détection, qu'elles reposent sur l'analyse d'artefacts statistiques, de caractéristiques visuelles, d'empreintes spécifiques aux architectures, ou sur des modèles d'apprentissage profond. Une attention particulière est portée aux recherches visant à renforcer la robustesse et la généralisation des détecteurs face aux attaques ou aux contenus de haute qualité.

Ce travail vise ainsi à articuler les enjeux cognitifs, sociétaux et techniques autour des GAN, en soulignant la nécessité d'outils adaptés pour préserver la confiance dans les images numériques.

Mots-clés : GAN, détection d'images synthétiques, artefacts visuels, empreintes numériques, forensique, deepfakes, perception humaine, robustesse, apprentissage automatique

Abstract

In recent years, generative models have drastically reshaped our relationship with digital imagery. Generative adversarial networks (GANs) can now produce images and videos so realistic that they are often indistinguishable from real content. While these advances enable new opportunities in many domains, they also raise major concerns about misinformation, visual forgery, and the erosion of trust in digital content.

This dissertation focuses on the detection of images and videos generated by GANs, but begins with an examination of human perceptual limitations and the societal implications raised by synthetic content. This initial framing helps contextualize the need for effective detection tools and responsible oversight.

We then provide a brief overview of visual manipulations before the rise of GANs, and introduce the core concepts of modern generative models. The main part of this work is devoted to the study of detection methods, including approaches based on statistical artifacts, visual anomalies, GAN-specific fingerprints, and deep learning models. We also review recent efforts aimed at improving detector robustness and generalization against attacks and high-quality generated content.

This work aims to articulate cognitive, social, and technical challenges related to GANs, emphasizing the need for adapted tools to help preserve trust in digital imagery.

Keywords: GAN, synthetic image detection, visual artifacts, fingerprints, forensics, deepfakes, human perception, robustness, machine learning

Indice

Table des figures	ix
Liste des tableaux	xv
Glossary	xvii
Liste des Acronymes	xxi
1 Introduction	1
2 Perception humaine et limites cognitives	3
2.1 Limites humaines dans la détection d'images falsifiées	3
2.2 Facteurs influents et biais cognitifs	4
2.3 Exemples visuels et types de falsification	4
2.4 Pourquoi l'aide humaine ne suffit plus	5
3 Perspectives, risques et enjeux pour la société	7
3.1 Évolution rapide des modèles génératifs	7
3.2 Défis techniques pour les détecteurs	8
3.3 Enjeux juridiques, réglementaires et éthiques	8
3.4 Traçabilité numérique et approches hybrides	8
4 Falsifications visuelles avant l'ère des GAN : typologie et méthodes de détection classiques	9
4.1 Typologie des falsifications visuelles historiques	9
4.1.1 Retouches numériques	9
4.1.2 Compositing et falsification contextuelle	10
4.1.3 Morphing facial et attaques biométriques	12
4.2 Enjeux de détection avant les GAN	14
4.2.1 Applications : biométrie, journalisme, documents officiels . .	14
4.3 Méthodes de détection non basées sur le machine learning	15
4.3.1 Détection du morphing classique	15
4.3.2 Analyse d'histogrammes	16
4.3.3 Compression JPEG et artefacts visibles	20
4.3.4 Analyse de bruit ou de duplications	23

4.4	Limites face à l'augmentation de la complexité des falsifications	25
5	Modèles génératifs : principes et évolution avant les GAN	27
5.1	Modèles génératifs vs discriminatifs : définitions	27
5.2	Modèles classiques : auto-encodeurs, VAE, modèles de Markov	29
5.2.1	Auto-encodeurs (AE)	29
5.2.2	Auto-encodeurs variationnels (VAE)	30
5.2.3	Modèles de Markov et chaînes latentes	31
5.3	Premières tentatives de génération d'images	32
5.4	Limites des approches pré-GAN	34
5.4.1	Expressivité restreinte des modèles probabilistes explicites . .	34
5.4.2	Qualité visuelle insuffisante	34
5.4.3	Difficultés d'entraînement et de convergence	35
5.4.4	Absence de critère perceptuel ou discriminatif	35
6	Les GAN : fondements, évolutions et capacités de génération	37
6.1	Principes fondamentaux des GAN	37
6.1.1	Principe du jeu adversarial	37
6.1.2	Formulation mathématique	38
6.1.3	Défis pratiques de l'entraînement	39
6.2	Architecture de base et variantes majeures	41
6.2.1	DCGAN : vers une architecture convolutive stable	42
Principes architecturaux	42	
Représentations apprises et applications	43	
Limites et apports	45	
6.2.2	ProGAN : génération progressive pour stabiliser la montée en résolution	46
Principe de la croissance progressive	47	
Bénéfices et apports techniques	47	
Résultats sur CelebA-HQ	48	
Limites et héritage	48	
6.2.3	StyleGAN et StyleGAN2 : séparation des styles et amélioration de la fidélité visuelle	49
Architecture : séparation explicite des styles	49	
Contrôle multi-échelle et mélange de styles	50	
Qualité visuelle et désentrelacement	50	
Améliorations apportées par StyleGAN2	51	
Impact et postérité	52	
6.3	Capacités de génération et détournements frauduleux	53

7 Méthodes de détection des images générées par GAN	55
7.1 Méthodes basées sur des descripteurs et artefacts visuels	55
7.1.1 Analyse de matrices de co-occurrence	55
Principe des matrices de co-occurrence	56
Architecture de détection proposée	56
Évaluation et robustesse	57
7.1.2 Analyse colorimétrique	58
Formation des couleurs : une analogie défaillante	59
Fréquence des pixels saturés : effet de la normalisation	60
Deux détecteurs complémentaires	61
Résultats expérimentaux	61
Discussion	62
7.1.3 Simulation d'artefacts	62
Artéfacts typiques et motivation de la simulation	62
Analyse fréquentielle des artefacts GAN	63
Réseau de simulation d'artefacts GAN	64
Apprentissage et détection supervisée	65
Résultats expérimentaux et robustesse	66
7.2 Méthodes fondées sur l'apprentissage automatique	67
7.2.1 Détection par CNN et SVM	67
Du descripteur SRM au CNN contraint	67
Comparaison aux approches SVM classiques	68
Empreintes résiduelles et identification de source	69
Discussion	69
7.2.2 Méthodes par empreinte GAN (fingerprints)	70
7.2.3 Analyse des points de repère faciaux	74
7.3 Amélioration de la robustesse et généralisation	76
7.3.1 Détection multi-indices	76
Robustesse à la compression sur les réseaux sociaux	78
7.3.2 Évaluation sur des images de haute qualité	79
7.4 Enjeux de résistance aux contre-attaques	80
7.4.1 Effets des post-traitements standards	80
7.4.2 Perturbations adversariales ciblées	81
7.4.3 Limites et perspectives	82
8 Conclusion	83
Références	85

Table des figures

2.1	Exemples d'images vraies et falsifiées avec superposition des clics utilisateurs (extrait de [1]).	5
4.1	Exemples d'images falsifiées et de fausses affirmations.	11
4.2	Exemple de morphing facial entre deux individus produisant une image hybride visuellement réaliste [2].	12
4.3	Exemple d'images créées automatiquement par morphing [2].	14
4.4	Exemples d'images CGG (générées par ordinateur) et PI (photographiques), accompagnées de la comparaison de leurs histogrammes de différences horizontales. Tiré de la Fig. 2 dans [3].	19
4.5	Effet d'une double compression JPEG sur l'histogramme du coefficient DCT (1, 2). Les pics périodiques témoignent d'une recompres- sion avec des facteurs de qualité différents [4].	21
4.6	Histogrammes des coefficients DCT (2, 1) dans une image originale, une image modifiée par F5 (stego), et l'estimation de l'histogramme réel. L'insertion modifie de manière systématique la répartition [4]. .	22
4.7	Résultats de détection de duplications issues de Fig. 2 dans [5]. Chaque ligne présente une image originale, son altération par copie, puis la détection des régions dupliquées.	24
5.1	Différences entre un modèle discriminatif ($P(y x)$) et un modèle gé- nératif ($P(x, y)$), dans une tâche de classification de chiffres manuscrits.	28
5.2	Schéma du processus de transfert de style neuronal. L'image d'entrée (en bas à gauche) est traitée par un réseau convolutionnel pour en extraire les représentations de contenu, tandis que l'image de style (en haut à gauche) fournit des représentations stylistiques. Les couches du réseau sont exploitées pour recomposer une nouvelle image combinant contenu et style à différents niveaux d'abstraction [6].	33
5.3	Séquences d'images générées par un RBM entraîné sur MNIST. Entre chaque paire d'images, une étape de Gibbs sampling est effectuée ($v \rightarrow$ $h \rightarrow v$). La première ligne illustre une transformation progressive d'un chiffre 8, la seconde une transition du chiffre 9 vers 7. [7].	34

6.1	Illustration du mécanisme d'apprentissage des GANs : au fil de l'entraînement, le générateur améliore ses productions tandis que le discriminateur affine sa frontière de décision, jusqu'à ce que les distributions synthétique et réelle deviennent indiscernables. [8]	40
6.2	Architecture du générateur DCGAN utilisé pour modéliser les scènes du jeu de données LSUN. Le vecteur latent est projeté, puis transformé par une série de convolutions transposées en une image 64×64 sans fully connected layer [9].	43
6.3	Visualisation guidée des activations de couches du discriminateur entraîné sur LSUN. Certaines activations sont spécifiques à des éléments visuels (lits, fenêtres) [9].	43
6.4	Interpolation entre vecteurs latents dans l'espace Z : les transitions entre chambres générées sont progressives et cohérentes [9].	44
6.5	Exemple d'arithmétique vectorielle dans l'espace latent des DCGAN. Des opérations linéaires sur les vecteurs latents permettent de manipuler des attributs visuels spécifiques [9].	45
6.6	Comparaison entre des échantillons générés par un GAN classique (milieu) et par un DCGAN• (droite), sur le dataset MNIST. Les images réelles sont affichées à gauche. [9]	46
6.7	Principe de la croissance progressive dans ProGAN : l'entraînement commence avec une faible résolution (4×4) à la fois pour le générateur (G) et le discriminateur (D). Des couches supplémentaires sont ensuite ajoutées progressivement, doublant à chaque étape la résolution spatiale des images synthétisées. Ce processus favorise une génération stable à haute résolution [10].	47
6.8	Exemples de visages synthétiques générés par ProGAN sur le dataset CelebA-HQ, à une résolution de 1024×1024 pixels. Cette qualité visuelle est atteinte grâce à l'apprentissage progressif, qui permet au modèle de capturer d'abord la structure globale avant de raffiner les détails fins [10].	48
6.9	Architecture du générateur de StyleGAN : un vecteur latent z est d'abord mappé vers un espace intermédiaire w , qui contrôle chaque couche via une normalisation adaptative. Un bruit stochastique est également injecté à chaque résolution pour enrichir les détails aléatoires [11].	49
6.10	Illustration du <i>style mixing</i> dans StyleGAN : chaque colonne résulte d'un mélange entre deux styles latents A et B à différentes résolutions (coarse, middle, fine). On observe que les couches profondes contrôlent la structure globale (pose, lunettes), tandis que les couches superficielles influencent les détails fins (peau, cheveux) [11].	50

6.11	Artefacts visibles dans StyleGAN, causés par la normalisation AdaIN. Ces distorsions, souvent en forme de bulles, apparaissent dès les premières couches de convolution (résolution 64×64) et sont présentes dans toutes les feature maps internes [12].	51
6.12	Suppression des artefacts dans StyleGAN2 grâce à la technique de <i>weight demodulation</i> . Les activations deviennent plus homogènes et les images synthétisées présentent une qualité visuelle plus régulière [12].	52
7.1	Exemples d'images générées par CycleGAN et StarGAN, extraits de [13].	56
7.2	Schéma de l'architecture complète proposée par Nataraj et al., combinant matrices de co-occurrence RGB et CNN [13].	57
7.3	Architecture simplifiée du générateur (tirée de [14]). La conversion des couches profondes vers la sortie RGB est réalisée par convolution 1×1 sur les canaux.	59
7.4	Poids de projection utilisés par le générateur GAN pour produire les canaux R, G, B à partir de 16 couches internes. La forte redondance et corrélation entre les trois courbes contraste avec les profils spectraux d'un capteur réel.	59
7.5	Comparaison visuelle et histogrammes : les images GAN (centre) tendent à éviter la saturation, contrairement aux images réelles qui présentent des plages surexposées (droite) ou sous-exposées (gauche) [14].	60
7.6	Courbes ROC pour le détecteur basé sur la saturation : bons résultats sur GAN Crop, performance réduite sur GAN Full.	61
7.7	Courbes ROC du détecteur de chromaticité : faible pouvoir discriminant.	62
7.8	Exemple d'artefact typique introduit par les GAN : motif périodique de type « damier », causé par les opérations de déconvolution [15]. .	63
7.9	Spectre de fréquence moyen d'un grand ensemble d'images GAN. On observe des pics anormaux dans certaines bandes spécifiques, absents dans les images naturelles [15].	63
7.10	Comparaison entre le spectre fréquentiel d'une image GAN (gauche) et d'une image réelle (droite). Les images GAN présentent des structures périodiques fortement localisées, absentes dans les images naturelles [15].	64
7.11	Pipeline complet de la méthode : le simulateur applique des artefacts artificiels, et le détecteur est entraîné à distinguer images réelles, images simulées, et images GAN [15].	65

7.12 Analyse fréquentielle des artefacts simulés : les distorsions introduites reproduisent des schémas proches de ceux observés dans les GAN réels [15].	66
7.13 Conversion du pipeline SRM vers un CNN contraint avec couches convolutionnelles et moyenne spatiale [16].	68
7.14 Comparaison de la localisation des zones falsifiées pour deux types de manipulations : splicing flouté (en haut) et copy-move redimensionné (en bas). De gauche à droite : image originale, image falsifiée, carte de chaleur obtenue avec SRM+SVM, carte obtenue avec le CNN proposé [16].	69
7.15 Empreintes estimées avec $N = 2, 8, 32, 128, 512$ résidus pour CycleGAN (haut) et ProGAN (bas) [17].	70
7.16 Autocorrélation de l'empreinte estimée pour ProGAN ($N = 512$) [17].	71
7.17 Distribution des corrélations entre résidus d'image et empreintes GAN : forte distinction entre empreintes croisées et correspondantes [17]. . .	71
7.18 Matrice de corrélation moyenne entre empreintes et résidus sur 22 sources (20 GANs et 2 caméras) [17].	72
7.19 Matrice de confusion pour l'attribution de source (GANs et caméras) [17].	73
7.20 Utilisation du fingerprinting dans le Forensics GAN Challenge [17]. .	73
7.21 Exemples d'anomalies dans des visages synthétisés par PGGAN : asymétrie des yeux (a), décalage de la bouche (b), coin de l'oeil anormalement anguleux (c) [18].	74
7.22 Pipeline proposé : détection des landmarks, normalisation, vectorisation, puis classification via SVM [18].	74
7.23 Distributions des coordonnées normalisées des points de repère sur les jeux de données CelebA (réel) et PGGAN (faux) [18].	75
7.24 Impact de la taille des images sur l'AUROC obtenu avec l'approche SVM [18].	76
7.25 Architecture d'exécution du modèle hiérarchique proposé par Guarnera et al. [19]. Le processus de classification s'effectue par niveaux : réel vs IA, puis GAN vs modèle de diffusion, et enfin identification de l'architecture générative.	77
7.26 Évolution de la précision et de l'erreur pour chaque niveau de classification du pipeline hiérarchique, durant l'apprentissage et le test [19].	78
7.27 Exemples de visages générés par StyleGAN2, tous de résolution 1024×1024 . Leur réalisme pose un défi majeur pour la détection automatique [21].	79

7.28 Architecture du détecteur CNN proposé par Nowroozi et al., basé sur les cooccurrences spatiales et croisées entre canaux (modèle Cross-Co-Net) [21].	79
---	----

Liste des tableaux

2.1	Matrice de confusion des réponses humaines (d'après [1]).	4
2.2	Corrélation entre variables utilisateurs et performance globale (selon [1]).	4
7.1	Expérience de généralisation croisée entre jeux de données	57
7.2	Effet de la compression JPEG sur la performance de détection . . .	58
7.3	Comparaison des précisions de classification sur CycleGAN par catégorie	58
7.4	Comparaison des performances entre CNN et SVM sur PGGAN/CelebA [18].	75
7.5	Robustesse comparée des modèles Cross-Co-Net et Co-Net sous divers post-traitements (StyleGAN2).	81

Glossaire

AdaIN

Technique de normalisation utilisée dans les réseaux neuronaux, notamment pour le transfert de style. Elle consiste à adapter les statistiques (moyenne et écart-type) d'un feature map cible selon des paramètres appris, afin de transférer une information de style à différentes échelles de l'image. Utilisée par StyleGAN pour moduler les caractéristiques générées à chaque niveau du réseau

blocking artifacts

artéfacts visuels en forme de blocs de 8×8 pixels résultant de la compression JPEG, perceptibles aux frontières entre blocs lorsque le taux de compression est élevé

CelebA-HQ

Version haute résolution (*High Quality*) du jeu de données *CelebA*, contenant 30 000 images de visages humains à 1024×1024 pixels. Utilisé pour entraîner et évaluer des modèles de génération d'images faciales haute fidélité, notamment ProGAN et StyleGAN

compositing

procédé de manipulation visuelle qui consiste à combiner plusieurs images ou éléments d'images en une seule composition visuelle cohérente

copy-move

type de falsification visuelle où une région d'une image est copiée puis collée ailleurs dans la même image, souvent pour masquer un élément ou dupliquer un motif

DCT

Transformée en cosinus discrète, utilisée notamment dans la compression JPEG pour exprimer l'image dans le domaine fréquentiel

Deep Belief Network

Modèle de réseau de neurones probabiliste composé de plusieurs couches empilées de *Restricted Boltzmann Machines* (RBMs), où les couches supérieures sont entraînées de manière non supervisée et les couches inférieures peuvent être ajustées de manière supervisée. Les DBN visent à apprendre des représentations hiérarchiques des données.

deepfake

vidéo ou image truquée générée par des modèles d'intelligence artificielle, souvent utilisée pour imiter des visages ou des expressions de manière réaliste

denoising

tâche de reconstruction d'une donnée corrompue ou bruitée à partir d'une version altérée, souvent utilisée pour évaluer la robustesse des modèles génératifs

fact-checking

procédure de vérification d'une information, d'un contenu ou d'une image, souvent utilisée pour contrer la désinformation

FFHQ

Flickr-Faces-HQ, un jeu de données de 70 000 visages haute résolution (1024×1024) couvrant une plus grande diversité de tranches d'âge, d'expressions faciales, de poses, d'ethnies et d'accessoires que CelebA-HQ. Introduit pour entraîner StyleGAN

FID

Pour *Fréchet Inception Distance*. Mesure de similarité entre la distribution des caractéristiques d'un ensemble d'images générées et celles d'un ensemble d'images réelles. Elle repose sur l'extraction de représentations par un réseau Inception, puis le calcul d'une distance de Fréchet entre les moyennes et matrices de covariance des deux ensembles. Une valeur plus faible indique une meilleure qualité et diversité des échantillons générés

forensique

terme issu du mot anglais forensics, désignant l'ensemble des méthodes d'analyse visant à authentifier ou vérifier l'origine d'un contenu numérique à des fins d'enquête ou de preuve

GAN

modèle d'apprentissage automatique composé d'un générateur et d'un discriminateur qui s'affrontent pour produire des données synthétiques réalistes.

GAN

Generative Adversarial Network, un réseau génératif antagoniste composé d'un générateur et d'un discriminateur en compétition dans un jeu à somme nulle

JSD

Divergence de Jensen-Shannon, mesure symétrique utilisée pour évaluer la similarité entre deux distributions de probabilité

matrice de co-occurrence

Représentation statistique de la fréquence conjointe d'apparition de paires de valeurs de pixels à des positions relatives données dans une image. Utilisée

notamment en stéganalyse et en traitement du signal pour détecter des régularités locales

minimax

Problème d'optimisation à deux joueurs dans lequel un acteur minimise une fonction tandis que l'autre la maximise

mode collapse

phénomène rencontré lors de l'entraînement des GANs où le générateur produit une faible diversité d'échantillons, ignorant certaines parties de la distribution réelle

modèle discriminatif

modèle d'apprentissage supervisé qui estime directement la probabilité conditionnelle $P(y | x)$ pour prédire une sortie y à partir d'une entrée x

modèle génératif

modèle qui cherche à approximer la distribution conjointe $P(x, y)$ ou la distribution marginale $P(x)$ afin de générer de nouvelles données réalistes

morphing

technique de transformation visuelle consistant à fusionner deux images, souvent deux visages, pour en créer une nouvelle contenant des caractéristiques des deux sources

pixel-wise normalization

Technique consistant à normaliser, pour chaque position spatiale (x, y) , le vecteur formé par les activations sur l'ensemble des feature maps, de manière à lui donner une norme unitaire. Utilisée dans le générateur de ProGAN pour stabiliser l'entraînement et éviter l'escalade des activations

réseau convolutif

Type de réseau de neurones conçu pour le traitement de données structurées en grille, notamment les images. Il repose sur l'application de filtres (ou noyaux) qui extraient automatiquement des motifs visuels à différentes échelles, tout en réduisant la dimensionnalité grâce à des opérations de convolution, de pooling et d'activation

tampering

terme anglais désignant toute forme de manipulation ou d'altération intentionnelle d'un contenu visuel dans le but de tromper ou de modifier la perception du spectateur

Liste des Acronymes

AE	auto-encodeur
C2PA	Coalition for Content Provenance and Authenticity
DCGAN	<i>Deep Convolutional Generative Adversarial Network</i>
HMM	<i>Hidden Markov Model</i>
PRNU	<i>Photo Response Non-Uniformity</i>
ProGAN	<i>Progressive Growing of GANs</i>
RBF	Radial Basis Function
RBM	<i>Restricted Boltzmann Machine</i>
SPAM	Subtractive Pixel Adjacency Matrix
SRM	Spatial Rich Model
StyleGAN	<i>Style-based Generator Architecture for GANs</i>
SVM	<i>Support Vector Machine</i>
VAE	auto-encodeur variationnel

Chapitre 1

Introduction

L'intelligence artificielle a profondément transformé notre manière de produire, percevoir et interpréter les images. Ces dernières années, des modèles comme les GAN — pour *Generative Adversarial Networks* — ont permis de générer des contenus visuels d'un réalisme saisissant. Introduits par Ian Goodfellow et al. [8], ces réseaux reposent sur une confrontation entre deux entités : un générateur, chargé de produire des images, et un discriminateur, chargé de les distinguer des images réelles. Ce processus compétitif améliore progressivement la qualité des images synthétiques, au point qu'elles deviennent parfois indiscernables pour l'œil humain.

À l'origine principalement développée dans un cadre expérimental, cette technologie s'est rapidement répandue au-delà de la recherche. Elle est aujourd'hui utilisée dans des domaines très variés : de la création artistique à la recherche médicale, en passant par le divertissement ou la publicité. Mais cette démocratisation s'accompagne aussi de dérives. Les deepfakes en sont un exemple frappant : ces vidéos truquées, souvent construites à partir de visages synthétisés, peuvent tromper l'opinion publique, servir des campagnes de désinformation ou encore nuire à la réputation d'une personne en l'associant à des contenus fabriqués de toutes pièces [22].

Dans ce mémoire, nous cherchons à explorer une question centrale : comment détecter les images et vidéos créées par des GANs, et quelles stratégies peut-on mettre en œuvre pour s'en protéger ? Au-delà des aspects strictement techniques, cette question soulève aussi des enjeux d'ordre éthique, social et juridique. Il ne s'agit pas

seulement d'arriver à repérer une image artificielle, mais aussi de réfléchir à ce que cela change dans notre rapport à l'image, et plus largement à la confiance que l'on accorde aux contenus numériques.

Nous avons choisi de débuter ce travail par une analyse des limites humaines face aux images synthétiques, ainsi que par une réflexion sur les risques sociaux associés. En interrogeant d'abord notre capacité à percevoir la vérité visuelle et à faire confiance à l'image, nous posons les bases d'un questionnement plus large sur la nécessité de méthodes de détection efficaces.

Nous reviendrons ensuite sur les falsifications visuelles antérieures à l'émergence des GAN. Bien avant l'arrivée de l'apprentissage automatique, des techniques comme la retouche, le compositing ou le morphing étaient déjà utilisées, que ce soit pour des objectifs artistiques ou pour tromper [23, 24, 2]. Nous proposerons également une mise en contexte des modèles génératifs afin de poser les bases nécessaires à la compréhension des GANs [25].

Une part centrale sera ensuite consacrée à l'étude des méthodes de détection. De nombreuses recherches ont montré que les GANs, malgré leur efficacité, laissent parfois des traces subtiles dans les images : motifs récurrents, anomalies colorimétriques, co-occurrences suspectes [14, 13]. Certaines approches exploitent même des « empreintes numériques » propres aux architectures de génération [17]. Nous analyserons en détail ces différentes familles de méthodes, en distinguant les approches sans apprentissage, celles basées sur l'apprentissage automatique, ainsi que les recherches récentes portant sur la robustesse et la généralisation des détecteurs.

À travers ce mémoire, notre objectif est d'apporter une contribution à la fois technique et réflexive sur un phénomène qui prend de l'ampleur. Plutôt que de nous limiter à l'analyse des GANs eux-mêmes, nous chercherons à articuler une compréhension globale : des limites perceptives aux méthodes de génération, jusqu'aux techniques de détection et à leurs implications concrètes pour la société. En croissant les approches, nous espérons offrir une lecture plus nuancée des risques — mais aussi des pistes de protection — face à la diffusion de contenus qui nécessitent des méthodes de fact-checking.

Chapitre 2

Perception humaine et limites cognitives

2.1 Limites humaines dans la détection d'images falsifiées

L'un des fondements de la lutte contre les images truquées repose sur la capacité de l'œil humain à identifier les incohérences visuelles. Or, cette aptitude, souvent surestimée, se révèle très limitée en pratique. L'étude menée par [1] constitue un repère important à ce sujet. Elle évalue, à travers une expérimentation à grande échelle, la capacité de 393 participants à détecter des falsifications dans 177 images issues de bases publiques forensiques.

Chaque participant devait, pour chaque image, indiquer si celle-ci avait été modifiée, et, le cas échéant, désigner une zone suspecte par un clic. Cette exigence d'argumentation visuelle permettait de distinguer les vraies détections des simples intuitions chanceuses. Les résultats sont frappants : sur plus de 17 000 réponses, l'exactitude globale des participants atteint seulement 57,5 %, et la sensibilité c'est-à-dire la capacité à repérer les images réellement falsifiées, ne dépasse pas 46,5 %.

Ces résultats suggèrent une tendance conservatrice : en cas de doute, les utilisateurs préfèrent répondre que l'image est authentique. Ce comportement est d'autant plus

	Image réelle (T)	Image falsifiée (F)
Réponse T (non falsifiée)	5 520 (T :T)	5 038 (F :T + F :Fi)
Réponse F (falsifiée)	2 271 (T :F)	4 379 (F :Fv)

TABLE 2.1 – Matrice de confusion des réponses humaines (d'après [1]).

problématique que certaines manipulations sont discrètes, comme dans le cas d'un simple effacement d'objet.

2.2 Facteurs influents et biais cognitifs

L'étude explore aussi les variables influençant la capacité de détection. Si l'on pourrait s'attendre à une influence significative de l'âge, du niveau d'études ou de l'expérience en traitement d'image, les corrélations sont en réalité faibles. La seule variable de fond légèrement associée à une meilleure performance est l'expérience déclarée (amateur ou professionnel), avec une corrélation de $\rho = 0.122$, $p = 0.015$.

Variable	ρ	<i>p</i> -valeur
Âge	-0.146	0.003
Éducation	-0.035	0.479
Expérience	0.122	0.015
Confiance déclarée	0.143	0.004
Temps avant indice	0.033	0.503
Pleine résolution ouverte	0.116	0.021

TABLE 2.2 – Corrélation entre variables utilisateurs et performance globale (selon [1]).

Du côté comportemental, les utilisateurs qui consultent les images en haute résolution ou qui se déclarent plus confiants dans leurs réponses ont de meilleures performances. À l'inverse, ceux qui passent beaucoup de temps à analyser une image ou demandent des indices sont souvent moins performants par effet de sur-analyse.

2.3 Exemples visuels et types de falsification

Un des apports intéressants de l'étude réside dans la comparaison des types de manipulation. Les images modifiées par effacement sont celles que les participants détectent le moins (38,5 % de réussite), contre 46,9 % pour les manipulations par copie, et 59,4 % pour le splicing. Cette hiérarchie s'explique par le fait que l'effacement ne laisse souvent aucun artefact visible, tandis que le splicing introduit des éléments pouvant sembler incongrus.

La figure suivante illustre quelques cas où les utilisateurs ont été attirés par des zones saillantes, même lorsque ce n'était pas là que se trouvait la modification :

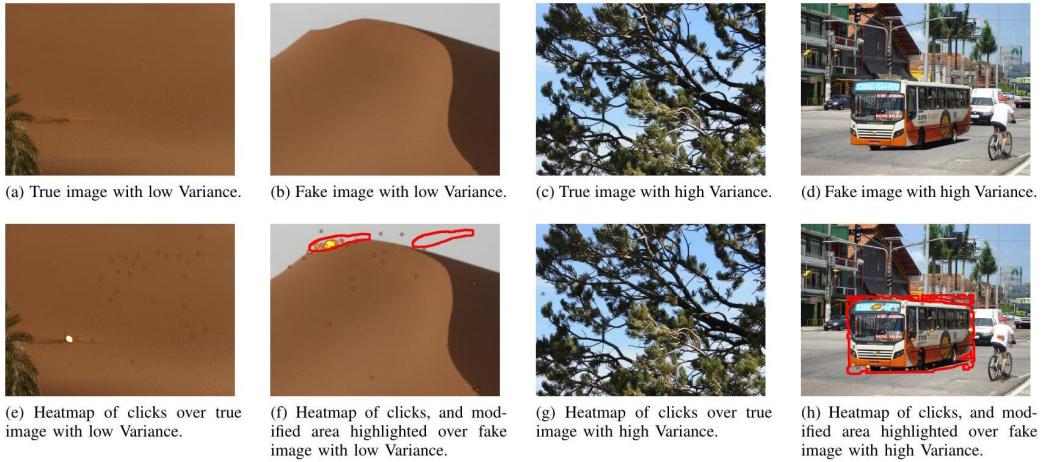


FIGURE 2.1 – Exemples d’images vraies et falsifiées avec superposition des clics utilisateurs (extrait de [1]).

Cette visualisation montre que les utilisateurs sont fortement influencés par des éléments visuellement saillants, même lorsqu’ils ne sont pas falsifiés – un phénomène amplifié par la complexité visuelle locale (variance).

2.4 Pourquoi l'aide humaine ne suffit plus

Ces résultats, obtenus avec des falsifications relativement classiques, sont déjà alarmants. Or, les GAN modernes produisent des images synthétiques encore plus difficiles à détecter. Contrairement aux collages visibles ou aux retouches approximatives, les deepfakes générés par GAN ne comportent souvent aucun artefact perceptible. La probabilité qu'un utilisateur puisse les détecter spontanément devient donc très faible.

Dans ce contexte, l'étude de [1] renforce l'idée que l'humain seul ne suffit pas. Elle justifie pleinement la mise en place d'outils de détection assistée. Ceux-ci doivent pouvoir repérer des indices faibles, imperceptibles à l'oeil nu (artefacts colorimétriques, motifs de compression, empreintes statistiques), et proposer des aides visuelles intelligentes, comme des cartes de probabilité ou des régions suspectes mises en évidence.

L'étude proposait un premier pas dans cette direction avec un système d'« indice visuel » : masquer la moitié non falsifiée d'une image pour concentrer l'attention sur l'autre moitié. Si ce système rudimentaire a pu aider certains participants, il a aussi parfois amplifié leurs biais. Cela montre que l'assistance visuelle doit être conçue

avec prudence : elle ne doit pas induire en erreur, mais orienter sans suggérer de conclusion automatique.

Les performances humaines, même dans un cadre contrôlé, sont insuffisantes pour faire face à la sophistication croissante des images truquées. L'étude de [1], bien qu'antérieure à l'essor des GAN, en fournit une démonstration quantitative claire. Elle légitime l'approche que nous adoptons dans ce mémoire : concevoir, évaluer et renforcer des méthodes de détection robustes, capables de combler les lacunes de la perception humaine à l'ère des images générées artificiellement.

Chapitre 3

Perspectives, risques et enjeux pour la société

L'essor fulgurant des modèles génératifs, en particulier ceux fondés sur les GAN, suscite des interrogations croissantes quant à leur impact sur nos sociétés. Ce chapitre propose une réflexion synthétique sur les perspectives qu'ouvrent ces technologies, tout en soulignant les défis techniques, éthiques et juridiques qui les accompagnent.

3.1 Évolution rapide des modèles génératifs

Depuis les premiers GAN de Goodfellow et al. [8], les progrès ont été saisisants. Des architectures comme StyleGAN2 ou GANverse3D permettent désormais de générer non seulement des visages d'un réalisme troublant, mais aussi des vidéos cohérentes, des objets en trois dimensions ou encore des scènes entières contrôlables par description textuelle. L'arrivée de modèles multimodaux, capables de produire des contenus à partir d'un simple prompt, pousse les frontières de la création artificielle encore plus loin. À mesure que ces modèles gagnent en résolution, en fidélité et en contrôle, la distinction entre le réel et le synthétique devient toujours plus floue.

3.2 Défis techniques pour les détecteurs

Face à cette montée en puissance, les outils de détection doivent sans cesse s'adapter. Chaque avancée dans la génération s'accompagne de nouveaux artefacts, mais aussi de stratégies pour les dissimuler. L'apprentissage profond a permis des progrès notables dans la détection, mais ceux-ci restent fragiles face à la généralisation, aux attaques adversariales, ou encore aux post-traitements comme la compression. Cette course entre génération et détection semble devoir se poursuivre, avec un équilibre sans cesse remis en question.

3.3 Enjeux juridiques, réglementaires et éthiques

Au-delà des considérations techniques, les usages de ces technologies soulèvent des questions fondamentales. Qui est responsable lorsqu'un deepfake est utilisé à des fins malveillantes ? Comment protéger les droits des individus dont l'image est exploitée sans leur consentement ? Les régulations en cours, notamment au sein de l'Union européenne, tentent d'encadrer ces pratiques, mais peinent à suivre le rythme de l'innovation. Dans le monde créatif également, des tensions émergent : un contenu généré par une intelligence artificielle peut-il être protégé par le droit d'auteur ? À qui revient la paternité de l'œuvre ? Ces débats, loin d'être tranchés, interrogent nos conceptions mêmes de l'originalité et de la création.

3.4 Traçabilité numérique et approches hybrides

Pour répondre à ces enjeux, plusieurs pistes sont explorées. L'une d'elles consiste à renforcer la traçabilité des contenus, en intégrant des signatures numériques, des empreintes persistantes ou des métadonnées vérifiables. Des consortiums tels que la Coalition for Content Provenance and Authenticity (C2PA) développent déjà des standards en ce sens. Parallèlement, l'avenir de la détection pourrait reposer sur des approches hybrides, mêlant analyse automatisée, validation humaine et traçabilité embarquée. Dans un monde où les images de synthèse se multiplient, garantir l'authenticité et la provenance d'un contenu pourrait devenir un impératif, tant pour préserver la confiance du public que pour assurer un usage éthique de ces technologies.

Chapitre 4

Falsifications visuelles avant l’ère des GAN : typologie et méthodes de détection classiques

4.1 Typologie des falsifications visuelles historiques

4.1.1 Retouches numériques

Modifier des images à l'aide de logiciels comme Photoshop est une pratique qui remonte à bien avant l'arrivée des modèles génératifs. Ces outils, très largement diffusés depuis les années 1990, ont rendu possible toutes sortes de transformations visuelles — certaines bénignes, d'autres beaucoup plus problématiques. En quelques clics, il est devenu facile de supprimer un élément gênant sur une photo, d'en dupliquer un autre, ou de réajuster localement la lumière, les contrastes, ou encore la couleur d'une scène.

Ces retouches, souvent réalisées pour des raisons esthétiques ou commerciales, peuvent aussi servir à manipuler l'interprétation d'une image. Le plus souvent, on pense à des modifications visibles : un objet effacé, un arrière-plan flouté, un visage « amélioré ». Mais dans certains cas, le travail est plus fin, plus discret — et donc plus difficile à

repérer.

Ce type de falsification laisse pourtant des traces. Wu et al. [3] ont notamment observé que ces retouches altèrent certaines régularités statistiques propres aux images naturelles, en particulier dans la distribution des niveaux de luminance ou de couleur. L'analyse des histogrammes peut ainsi révéler des incohérences, même lorsque les modifications sont peu perceptibles à l'oeil nu. [3]

Ce qu'il faut retenir, c'est que ces pratiques se sont banalisées. Aujourd'hui, n'importe qui peut retoucher une image depuis son téléphone. Cela pose un vrai défi, surtout dans des domaines où la véracité d'un contenu visuel a une importance critique : presse, documents officiels, réseaux sociaux. Bien avant qu'on parle de deepfake ou de GAN, ces manipulations ont soulevé les premières questions sur la confiance que l'on peut accorder à une image.

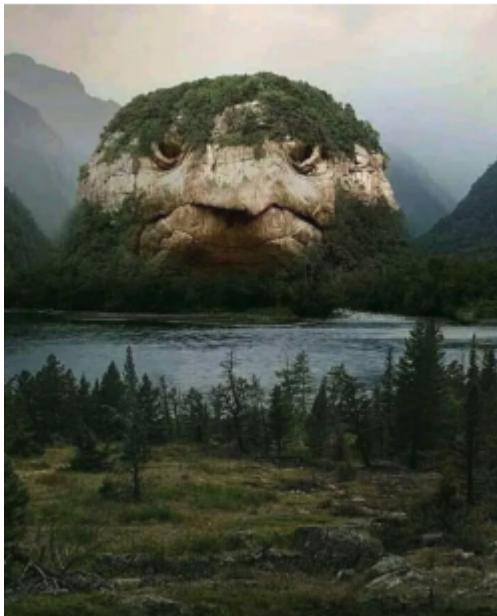
4.1.2 Compositing et falsification contextuelle

Une autre manière répandue de falsifier une image avant les techniques génératives modernes est d'assembler plusieurs éléments visuels issus de sources différentes. Ce procédé est le compositing, qui repose sur la superposition, l'incrustation ou la fusion d'objets, de visages ou de scènes entières pour produire une image cohérente en apparence, mais qui est entièrement recomposée.

Ce type de manipulation ne se limite pas à des modifications des pixels, il implique également un travail sur la perspective, la lumière, les ombres, ou encore la cohérence des couleurs entre les différentes parties assemblées. Il peut s'agir par exemple de remplacer le visage d'une personne sur une photo de groupe, de copier une silhouette d'une image à un autre ou encore d'intégrer un objet dans un décor [26].

Dans certains cas, il n'y a pas besoin de modifier techniquement l'image. Il suffit de changer le cadre ou le contexte dans lequel elle est présentée. Un recadrage stratégique peut éliminer un détail essentiel du jugement ou, au contraire, en exagérer un autre. Une légende mensongère peut détourner le sens initial d'une photographie authentique [26].

Ces falsifications dites *contextuelles* sont particulièrement redoutables, car elles exploitent les biais d'interprétation du spectateur. Même une image non modifiée, présentée dans un mauvais cadre narratif, peut induire en erreur l'observateur. Et contrairement aux retouches classiques, elles ne laissent pas de traces visibles sur le plan technique, ce qui rend leur détection d'autant plus délicate.



(a) Une photographie montre une montagne ressemblant à une tortue.



(b) Une photographie montre le président russe Vladimir Poutine tirant agressivement la cravate de l'ancien président américain Barack Obama.

FIGURE 4.1 – Exemples d’images falsifiées et de fausses affirmations.

Face à ce type de manipulation, les approches fondées uniquement sur des caractéristiques visuelles sont souvent insuffisantes. La compréhension du contexte, des métadonnées associées, ou de la cohérence entre image et texte devient alors cruciale. Des initiatives de fact-checking sont apparues ces dernières années, portées par des médias, des institutions ou encore certaines plateformes numériques. Ces dispositifs reposent sur des analyses techniques et de vérifications contextuelles où l'on effectue des comparaisons avec des images d'archives, le traçage de la source originale, ou encore un recouplement avec d'autres informations disponibles. Certains outils permettent également d'effectuer des recherches inversées d'images, afin d'identifier des duplications ou des réutilisations hors contexte [27].

Néanmoins, le fact-checking nécessite encore l'intervention humaine, qui peut être influencée ou biaisée. Zlatkova et al. [27] soulignent notamment que l'efficacité de ces vérifications sont limités avec des scènes complexes, dépend de la rapidité avec laquelle elles sont effectuées, ainsi que de la capacité du public à y accéder et à en tenir compte. Dans une société où l'information circule extrêmement vite, une image trompeuse peut être vue des millions de fois avant même d'être vérifiée. Cela souligne donc l'importance de combiner ces approches avec des techniques de détection automatisées et préventives pour limiter la diffusion de faux contenus avant qu'ils ne deviennent virales.

4.1.3 Morphing facial et attaques biométriques

Le morphing facial est un processus de falsification d'image qui fusionne deux visages pour en produire un nouveau, qui conserve les traits distinctifs des deux individus originaux [24].

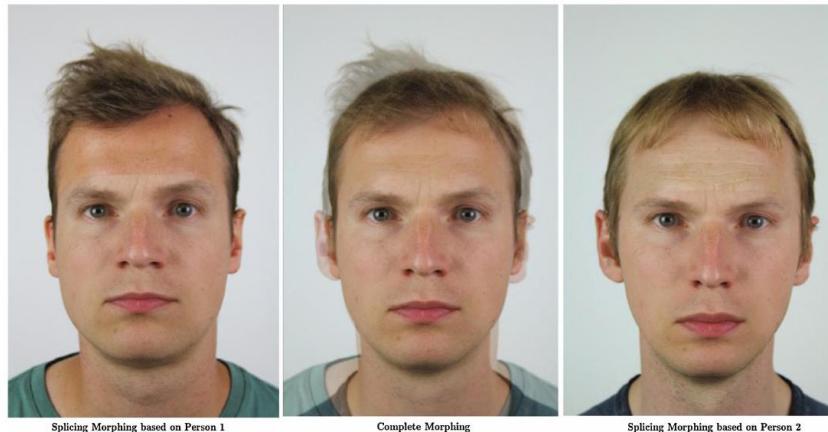


FIGURE 4.2 – Exemple de morphing facial entre deux individus produisant une image hybride visuellement réaliste [2].

Techniquement, le morphing facial repose sur une interpolation géométrique et photométrique entre deux visages sources. Le processus standard comprend deux étapes principales : le *warp*, qui déforme la géométrie des visages, et le *blend*, qui fusionne les valeurs de couleur. Cette méthode, décrite notamment par Wolberg [24], peut être modélisée de manière simple à l'aide d'interpolations linéaires.

Soient I_1 et I_2 deux images de visages, et $P_1 = \{(x_1^i, y_1^i)\}$, $P_2 = \{(x_2^i, y_2^i)\}$ les ensembles de points-clés détectés sur chaque visage (généralement 68 landmarks). Pour chaque point i , on définit le point intermédiaire au temps $t \in [0, 1]$ par interpolation linéaire :

$$(x_t^i, y_t^i) = (1 - t)(x_1^i, y_1^i) + t(x_2^i, y_2^i) \quad (4.1)$$

À partir de ces correspondances, on applique un *warp* sur les deux images vers cette géométrie intermédiaire, généralement en procédant par triangulation de Delaunay sur $P_t = \{(x_t^i, y_t^i)\}$. Une fois les deux images alignées géométriquement, on effectue une interpolation pixel à pixel (ou région par région) des valeurs colorimétriques pour obtenir l'image fusionnée :

$$I_t(x, y) = (1 - t) \cdot I_1(x_1, y_1) + t \cdot I_2(x_2, y_2) \quad (4.2)$$

où (x_1, y_1) et (x_2, y_2) sont les coordonnées des pixels correspondants dans I_1 et

I_2 obtenues par déformation inverse depuis le point (x, y) dans la morphologie intermédiaire.

Ce pipeline, bien qu'implémenté aujourd'hui dans des logiciels comme GIMP ou OpenCV, reste fondamentalement basé sur ces principes d'interpolation géométrique et colorimétrique.

Initialement développée à but artistique et de divertissement [24], cette méthode a rapidement soulevé des préoccupations en matière de sécurité biométrique, notamment pour l'authentification basée sur la reconnaissance faciale [2].

En termes d'usurpation d'identité, le morphing facial pose un risque majeur pour les contrôles où des documents d'identité sont demandés. Une fois ceux-ci obtenus avec une image morphée, il devient possible de passer les contrôles faciaux sans éveiller de soupçons [2].

La détection de telles manipulations est un défi important, car contrairement aux retouches ou au compositing où des anomalies visuelles sont souvent perceptibles, les images issues de morphing sont souvent très réalistes et conservent une structure cohérente. Plusieurs approches ont été proposées pour tenter de détecter les visages morphés. Certaines reposent sur l'observation d'artefacts après fusion entre deux visages. D'autres sur l'extraction de caractéristiques qui parfois permettent de repérer des incohérences statistiques présentes dans l'image [24, 2]. Il n'existe donc pas de solution unique, mais de telles données peuvent alerter sur l'authenticité d'une image.

La plupart des systèmes de reconnaissance faciale classiques restent très vulnérables à ce type de manipulation. Comme ils n'ont pas été conçus pour distinguer un visage authentique d'un visage synthétisé par morphing, ils peuvent valider un document falsifié sans s'en rendre compte [2]. Des pistes sont actuellement explorées, par exemple la comparaison entre plusieurs photos d'un même individu, ou l'analyse des textures très localisées du visage. Ces méthodes visent à identifier des anomalies qui, à l'œil nu, passeraient totalement inaperçues.

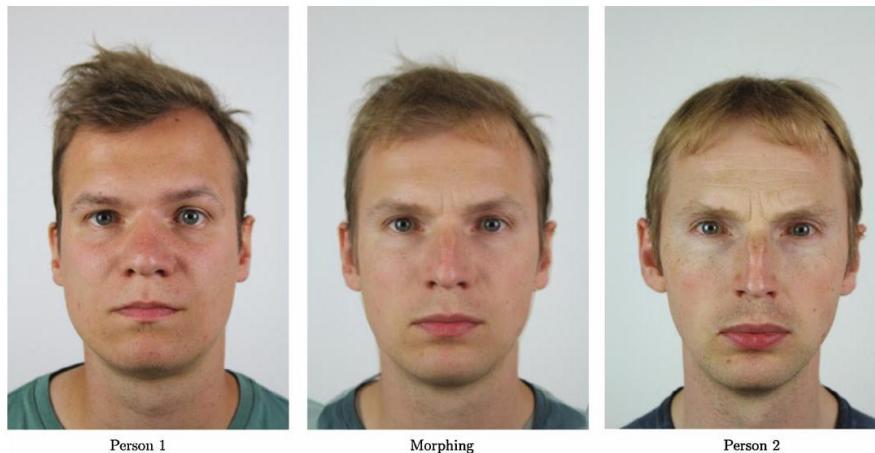


FIGURE 4.3 – Exemple d’images créées automatiquement par morphing [2].

Avec l’essor de cette technique, les enjeux pour la sécurité biométrique sont devenus plus visibles. De plus en plus de travaux cherchent à intégrer des solutions de détection directement dans les procédures administratives, notamment lors de l’émission de documents officiels. On observe aussi une tendance vers des prises de vue plus contrôlées, dans le but de rendre plus difficile l’utilisation de photos falsifiées. Le sujet reste ouvert, mais il est désormais pris au sérieux par les autorités concernées.

4.2 Enjeux de détection avant les GAN

4.2.1 Applications : biométrie, journalisme, documents officiels

Avant même l’essor des modèles génératifs, la détection de tampering posait déjà de sérieux défis dans plusieurs domaines sensibles. L’image, en tant que support d’information, de preuve ou d’identité, occupe une place centrale dans de nombreux processus de validation ou de diffusion. Lorsqu’elle est altérée, les conséquences peuvent être lourdes, en particulier dans des contextes où l’authenticité visuelle conditionne la confiance.

En biométrie, les images du visage sont devenues un identifiant courant utilisé dans les passeports, les cartes d’identité ou encore en reconnaissance faciale. Ces systèmes permettent de vérifier la correspondance entre un visage et une photo donnée mais sans toujours intégrer des dispositifs de vérification de l’authenticité des images elles-mêmes. Ainsi, une photo falsifiée — qu’elle soit retouchée, issue d’un morphing, ou modifiée par compositing — peut suffire à contourner les contrôles si elle est visuellement crédible [2]. Ce type de faille est d’autant plus dangereux qu’il touche à des enjeux de sécurité nationale, de lutte contre la fraude documentaire ou encore de

contrôle aux frontières.

Dans le champ journalistique, les images jouent un rôle crucial dans la représentation des faits. Une photographie retouchée ou sortie de son contexte peut facilement modifier la perception d'un événement, voire tromper le public. Ces formes de tampering sont particulièrement problématiques à l'ère des réseaux sociaux, où la vérification — fact-checking — est souvent négligée et où l'image circule bien plus vite que l'information elle-même [27]. Un exemple illustratif est présenté en figure 4.1, où l'image est détournée par une légende fallacieuse.

Au-delà de la biométrie, d'autres types de documents officiels peuvent aussi être affectés : justificatifs falsifiés, montages de documents administratifs, ou encore images utilisées dans des contextes juridiques. Une simple manipulation visuelle peut ainsi suffire à produire une fausse preuve, un faux relevé ou une fausse pièce d'identité. Ces usages, bien que souvent discrets, renforcent le besoin d'outils fiables de détection.

4.3 Méthodes de détection non basées sur le machine learning

4.3.1 Détection du morphing classique

Parmi les différentes formes de falsifications visuelles existantes, le morphing facial a représenté historiquement un cas d'étude emblématique, notamment en raison de ses implications critiques en biométrie. Avant même d'aborder des méthodes plus générales exploitant des caractéristiques globales ou spectrales, il est pertinent de rappeler les approches spécifiques qui ont été développées pour détecter ces falsifications bien particulières.

Le morphing facial, en tant que méthode de falsification visuelle, pose un défi particulier pour les systèmes de reconnaissance biométrique. Contrairement à des retouches simples ou à des duplications, cette technique génère une image réaliste qui partage des traits biométriques avec deux individus distincts. Une image ainsi produite peut tromper un système de vérification automatique en étant acceptée pour les deux visages originaux.

Avant l'introduction des modèles d'apprentissage profond, les méthodes classiques de détection du morphing s'appuyaient sur des approches statistiques et fréquentielles. Plusieurs travaux, comme ceux de Neubert et al. [2], ont étudié l'effet du morphing sur la qualité intrinsèque de l'image, observant qu'un visage morphé présente souvent une dégradation locale de la netteté ou une transition douce anormale

dans certaines régions du visage (yeux, nez, contour du visage).

Une méthode courante repose sur l’analyse des différences entre plusieurs images d’un même individu, par exemple lorsqu’une personne fournit plusieurs photos au fil du temps (données historiques, passeports précédents). Une image falsifiée par morphing tend à être moins cohérente avec le reste du jeu de données, que ce soit en termes de textures, de géométrie ou de distribution des couleurs.

D’autres approches exploitent les artefacts introduits lors de la fusion morphologique. En particulier, certaines méthodes analysent les incohérences dans les hautes fréquences (bordures mal alignées, textures floues localisées), ou la structure des contours du visage. Descripteurs comme les LBP (Local Binary Patterns), les HOG (Histogram of Oriented Gradients) ou des mesures de symétrie faciale ont été utilisés pour capturer ces anomalies.

Enfin, une piste explorée est celle de la compression : les visages morphés, du fait de leurs zones intermédiaires lissées, ont tendance à présenter des artefacts particuliers lors de la recompression JPEG. Des travaux ont ainsi montré que l’analyse d’artefacts de compression pouvait contribuer à détecter ces cas, en complément des approches précédentes.

Ces méthodes restent limitées en précision face à des morphings de haute qualité, mais elles ont constitué les premières tentatives systématiques pour identifier ce type de falsification sans apprentissage. Elles ont aussi jeté les bases des approches modernes fondées sur l’apprentissage machine supervisé ou non supervisé.

4.3.2 Analyse d’histogrammes

L’analyse par histogramme fait partie des approches les plus anciennes utilisées pour authentifier des images retouchées. Elle consiste à examiner comment se répartissent les niveaux de gris ou les couleurs dans une image. Lorsqu’une photo est prise de manière naturelle, sans manipulation, cette répartition — appelée distribution statistique — suit un certain équilibre. Elle tend à être continue, relativement lisse, et reflète les caractéristiques optiques du monde réel.

Mais dès qu’une modification est apportée à l’image, même minime, cette harmonie peut être perturbée. C’est cette déviation dans l’image que met en évidence l’analyse par histogramme, tandis qu’une modification artificielle introduit généralement des ruptures, des sauts ou des irrégularités dans cette distribution.

Wu et al. [3] ont proposé une méthode fondée sur l'extraction de caractéristiques statistiques directement issues des histogrammes des composantes RVB (Rouge, Vert, Bleu) d'une image. Leur objectif est de distinguer les images générées par ordinateur des photographies réelles, en supposant que la structure des histogrammes diffère significativement entre ces deux catégories. Plus précisément, ils calculent plusieurs mesures dérivées de l'histogramme, telles que la variance inter-bande, le degré de régularité locale, ou encore l'écart-type des pics dominants, pour capter des motifs typiques des images de synthèse ou modifiées.

La méthode proposée repose sur une idée simple mais efficace : plutôt que de transformer ou modéliser les histogrammes comme dans les approches classiques (via des moments, transformées de Fourier, etc.), les auteurs utilisent directement les bins les plus significatifs des histogrammes de différences comme vecteurs de caractéristiques.

Concrètement, ils construisent des *images de différences* en appliquant un filtre de dérivée première sur chaque canal couleur d'une image I . Pour une direction donnée i (par exemple horizontale), cela revient à calculer :

$$I_i = I * f_i \quad (4.3)$$

où f_i est un noyau de convolution, par exemple $f_h = (1, -1)$ pour la direction horizontale. L'image de différences I_i contient alors les écarts d'intensité entre pixels adjacents.

À partir de cette image, ils construisent un histogramme normalisé H défini comme suit :

$$H(n) = \frac{\#\{(x, y) \mid I_i(x, y) = n\}}{N}, \quad -255 \leq n \leq 255 \quad (4.4)$$

où N est le nombre total de pixels de l'image de différences, et $\#$ désigne la cardinalité de l'ensemble des pixels ayant une différence égale à n .

Afin de réduire la dimensionnalité tout en capturant la structure symétrique de l'histogramme (centrée autour de zéro), ils ne retiennent que $1 + k$ composantes de H agrégées comme suit :

$$H(0), \frac{H(1) + H(-1)}{2}, \dots, \frac{H(k) + H(-k)}{2} \quad (4.5)$$

Cette représentation compacte permet de saisir la concentration des variations d'intensité autour de la valeur nulle (ce qui reflète les régularités locales). Ce schéma est étendu à plusieurs directions et ordres de dérivation. En plus des différences horizontales (f_h), sont également considérées : - $f_v = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$ (verticale), - $f_d = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$

(diagonale), $-f_a = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$ (anti-diagonale),

ainsi que leurs combinaisons secondees $I_{i,j} = I * f_i * f_j$, avec $i, j \in \{h, v, d, a\}$, ce qui donne 14 images de différences (4 premières directions et 10 secondes-ordres distincts, en tenant compte de la symétrie $I_{i,j} = I_{j,i}$).

Pour chaque image de différences, on extrait les $1 + k$ composantes définies en 4.5, ce qui donne un total de $14(1 + k)$ caractéristiques par canal couleur. Les auteurs utilisent généralement $k = 7$, soit un vecteur de 112 dimensions par canal (et donc 336 si tous les canaux sont utilisés).

L'avantage majeur de cette méthode est sa simplicité computationnelle (aucune transformée complexe n'est nécessaire) et son efficacité : les expériences rapportées par Wu et al. [3] montrent qu'avec ces seules 112 caractéristiques, leur méthode atteint un taux de détection supérieur à 95% pour les images synthétiques, avec une aire sous la courbe ROC (AUC) proche de 0.99 (voir Tab1 et Fig5 de l'article).

Dans leur protocole expérimental, les auteurs analysent des histogramme de plus de 10000 images réparties entre vraies photographies et images générées. Le modèle repose sur des règles statistiques simples, comme le dépassement de certains seuils sur des indicateurs de texture ou de saturation. Par exemple, ils montrent que les images artificielles présentent souvent des pics plus marqués à des intensités spécifiques, dus à l'utilisation de couleurs standards ou de palettes limitées lors du rendu graphique.

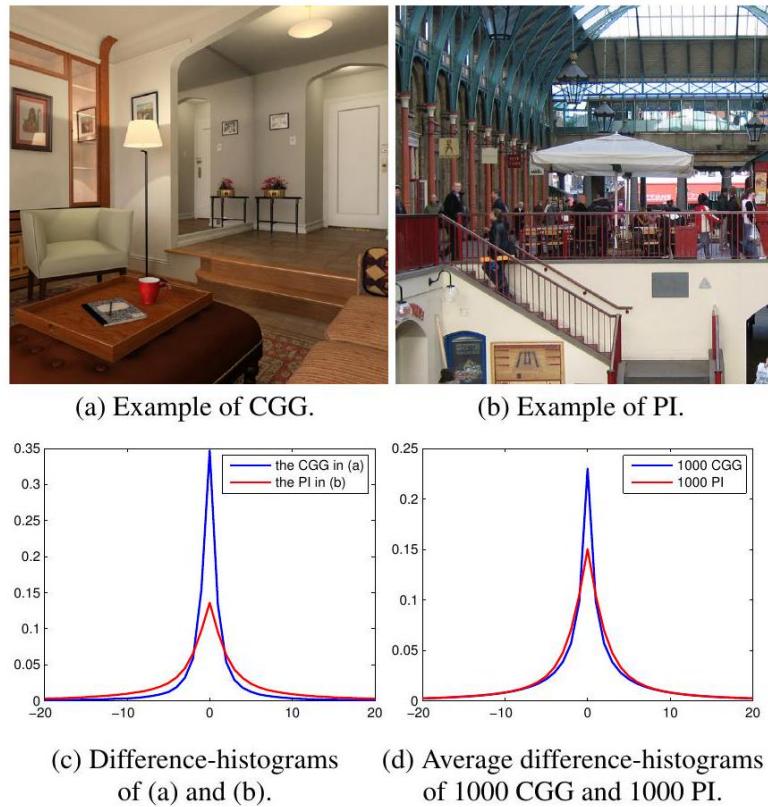


FIGURE 4.4 – Exemples d’images CGG (générées par ordinateur) et PI (photographiques), accompagnées de la comparaison de leurs histogrammes de différences horizontales. Tiré de la Fig. 2 dans [3].

Cette méthode présente deux avantages majeurs : elle est à la fois rapide (car elle repose sur des statistiques simples) et agnostique par rapport au type de falsification (aucune hypothèse sur la nature du tampering). Toutefois, elle souffre également de limites. En particulier, des falsifications sophistiquées peuvent préserver la structure globale de l’histogramme, notamment lorsqu’elles manipulent des régions de faible contraste ou exploitent des modèles photo-réalistes. De plus, certains traitements postérieurs à la falsification (floutage, compression, ou équilibrage des couleurs) peuvent masquer les irrégularités.

Bien que rudimentaire, l’analyse par histogramme reste une méthode pertinente dans un cadre exploratoire ou en complément d’autres approches forensique. Son intérêt réside dans sa simplicité, sa rapidité, et sa capacité à détecter des anomalies globales de distribution, notamment dans les cas de retouches non maîtrisées ou de génération graphique basique.

4.3.3 Compression JPEG et artefacts visibles

La compression JPEG est omniprésente dans les flux numériques d’images. Elle repose sur une série d’opérations visant à réduire la redondance visuelle, tout en maintenant une qualité perçue acceptable. Ce processus, bien qu’efficace, introduit également des artefacts caractéristiques, particulièrement visibles en cas de compression à faible qualité ou de recompression successive. Ces artefacts peuvent servir pour la détection de tampering.

Le pipeline de compression JPEG repose sur le découpage de l’image en blocs de 8×8 pixels sur lesquels est effectué une DCT [28, 29] :

$$F(u, v) = \frac{1}{4} \alpha(u) \alpha(v) \sum_{x=0}^7 \sum_{y=0}^7 f(x, y) \cos \left[\frac{(2x+1)u\pi}{16} \right] \cos \left[\frac{(2y+1)v\pi}{16} \right] \quad (4.6)$$

où $f(x, y)$ est l’intensité du pixel en position (x, y) , $F(u, v)$ le coefficient DCT obtenu, et $\alpha(u) = 1/\sqrt{2}$ si $u = 0$, $\alpha(u) = 1$ sinon. Cette transformation permet de concentrer l’information dans les basses fréquences.

Les coefficients $F(u, v)$ sont ensuite divisés par une matrice de quantification $Q(u, v)$ dépendant du facteur de qualité, et arrondis à l’entier le plus proche [28] :

$$\hat{F}(u, v) = \text{round} \left(\frac{F(u, v)}{Q(u, v)} \right) \quad (4.7)$$

Cette étape entraîne une perte d’information irréversible. Une fois l’image décompressée, des discontinuités peuvent apparaître notamment aux jonctions entre les blocs de 8×8 pixels. Ces irrégularités, appelées blocking artifacts, deviennent particulièrement visibles lorsque l’image a subi une forte compression.

L’image est ensuite reconstruite par l’application inverse de la DCT, en remultipliant les coefficients quantifiés par la matrice Q et en les transformant en blocs spatiaux :

$$f'(x, y) = \sum_{u=0}^7 \sum_{v=0}^7 \hat{F}(u, v) Q(u, v) \cdot \cos \left[\frac{(2x+1)u\pi}{16} \right] \cos \left[\frac{(2y+1)v\pi}{16} \right] \quad (4.8)$$

Cette reconstruction approximative peut introduire des erreurs visibles, notamment lorsque les coefficients DCT ont été fortement tronqués.

Dans un contexte forensique, plusieurs stratégies exploitent ces artefacts pour identifier des zones modifiées :

- la recherche de discontinuités anormales à l'intérieur de blocs homogènes [30],
- l'analyse de la périodicité des coefficients DCT quantifiés,
- la détection de double compression JPEG (lorsqu'une région retouchée a été recompressée localement).

Une image modifiée localement (ex. : insertion ou recadrage) perturbe la régularité des coefficients DCT, et ces anomalies sont détectables par des méthodes statistiques. Par exemple, une recompression locale engendre des pics périodiques caractéristiques dans les histogrammes :

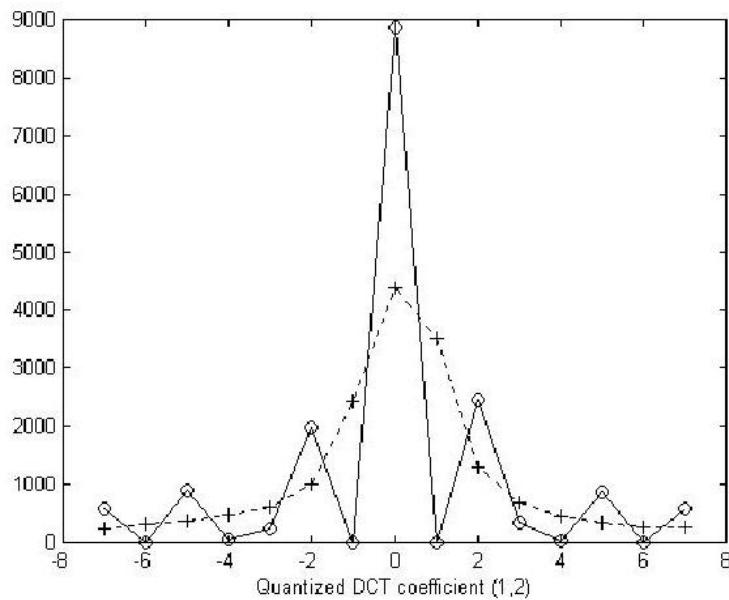


FIGURE 4.5 – Effet d'une double compression JPEG sur l'histogramme du coefficient DCT (1,2). Les pics périodiques témoignent d'une recompression avec des facteurs de qualité différents [4].

D'autres types de manipulations, comme l'insertion de message dans le domaine DCT (cas de la stéganographie), produisent également des altérations détectables dans la distribution des coefficients. C'est notamment le cas de l'algorithme F5, dont l'impact sur l'histogramme d'un coefficient (2, 1) est présenté ci-dessous :

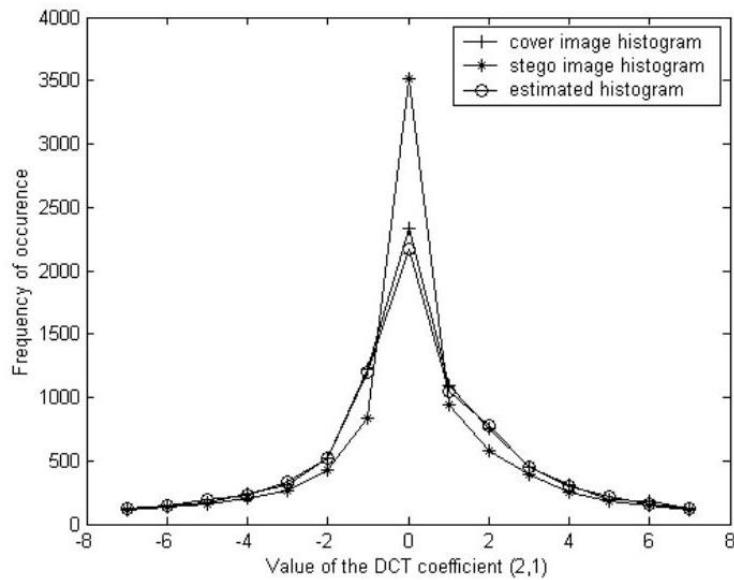


FIGURE 4.6 – Histogrammes des coefficients DCT (2,1) dans une image originale, une image modifiée par F5 (stego), et l'estimation de l'histogramme réel. L'insertion modifie de manière systématique la répartition [4].

Fridrich et al. [4] modélisent ces altérations par une équation linéaire entre histogrammes estimés et observés :

$$H_{kl}(d) = (1 - \beta) \cdot h_{kl}(d) + \beta \cdot h_{kl}(d + 1), \quad \text{pour } d > 0, \quad (4.9)$$

$$H_{kl}(0) = h_{kl}(0) + \beta \cdot h_{kl}(1), \quad (4.10)$$

où $H_{kl}(d)$ désigne l'histogramme du coefficient (k, l) de l'image suspecte, $h_{kl}(d)$ l'histogramme estimé du document original, et β la proportion de modifications.

La valeur de β peut être estimée en minimisant une erreur quadratique entre les histogrammes [4] :

$$\beta_{kl} = \arg \min_{\beta} \left[H_{kl}(0) - \hat{h}_{kl}(0) - \beta \hat{h}_{kl}(1) \right]^2 + \left[H_{kl}(1) - (1 - \beta) \hat{h}_{kl}(1) - \beta \hat{h}_{kl}(2) \right]^2 \quad (4.11)$$

ce qui conduit à l'expression fermée :

$$\beta_{kl} = \frac{\hat{h}_{kl}(1) \left(H_{kl}(0) - \hat{h}_{kl}(0) \right) + \left(H_{kl}(1) - \hat{h}_{kl}(1) \right) \left(\hat{h}_{kl}(2) - \hat{h}_{kl}(1) \right)}{\hat{h}_{kl}^2(1) + \left(\hat{h}_{kl}(2) - \hat{h}_{kl}(1) \right)^2} \quad (4.12)$$

L'estimation de β permet d'identifier la présence et l'intensité d'un contenu potentiellement inséré, même sans connaître la nature exacte de la falsification.

Plusieurs travaux ont proposé des approches efficaces fondées sur la distribution de ces coefficients, notamment à travers des modèles de Markov [31] ou des analyses spectrales. Ces méthodes sont non supervisées ou faiblement supervisées, et permettent une détection robuste sans apprentissage profond.

La compression JPEG introduit des artefacts exploitables pour la détection de falsifications visuelles. Leur analyse permet de révéler des zones modifiées, d'identifier des incohérences de structure, ou de détecter une recompression locale. Ces techniques s'intègrent naturellement dans les outils de forensique visuelle et constituent une brique fondamentale avant l'ère des méthodes basées sur l'apprentissage.

4.3.4 Analyse de bruit ou de duplications

Outre les artefacts de compression ou les altérations spectrales, les falsifications peuvent également perturber des propriétés plus subtiles d'une image, telles que son bruit résiduel ou sa structure auto-similaire. Ces indices, souvent imperceptibles à l'oeil nu, ont été exploités dans plusieurs travaux pour mettre en évidence des modifications locales, même en l'absence d'altérations visibles.

Analyse du bruit résiduel. Lorsqu'une image est acquise par un capteur numérique, elle contient du bruit électronique caractéristique du dispositif de capture. Ce bruit est généralement stable sur toute l'image, car il provient de l'optique, du capteur et du pipeline de traitement. Si une région de l'image a été copiée, insérée, ou manipulée à partir d'une autre source (ou même d'une autre partie de l'image), alors le profil de bruit local dans cette zone peut différer du reste. L'idée centrale de ces approches est donc d'extraire une carte du bruit local, par exemple en utilisant un filtre passe-haut, des résidus de prédiction, ou encore des modèles statistiques, puis de rechercher des incohérences spatiales [32].

Certaines méthodes vont plus loin en extrayant la signature de bruit *Photo Response Non-Uniformity* (PRNU) propre à un appareil photo et en la comparant aux différentes régions de l'image pour repérer des zones suspectes qui ne suivent pas le motif global attendu. Ces techniques sont particulièrement efficaces pour repérer des insertions d'éléments provenant d'autres sources.

Détection de duplications internes. Une autre forme courante de manipulation consiste à copier une zone d'une image pour la recoller ailleurs dans la même image – typiquement pour masquer un objet ou dupliquer un motif. Ce type de falsification, dit copy-move, laisse des empreintes exploitables, notamment en raison des petites différences de textures ou d'alignement, même après transformation géométrique.

Les techniques de détection de duplications reposent souvent sur un découpage de l'image en blocs superposés, sur lesquels sont extraites des signatures locales (par exemple via des descripteurs Zernike, DCT ou PCA), avant d'appliquer une recherche de correspondances entre blocs similaires [5]. Lorsque deux zones apparaissent trop semblables selon un certain seuil de distance, et qu'elles suivent une transformation géométrique cohérente (translation, rotation), cela constitue un indice fort de manipulation.

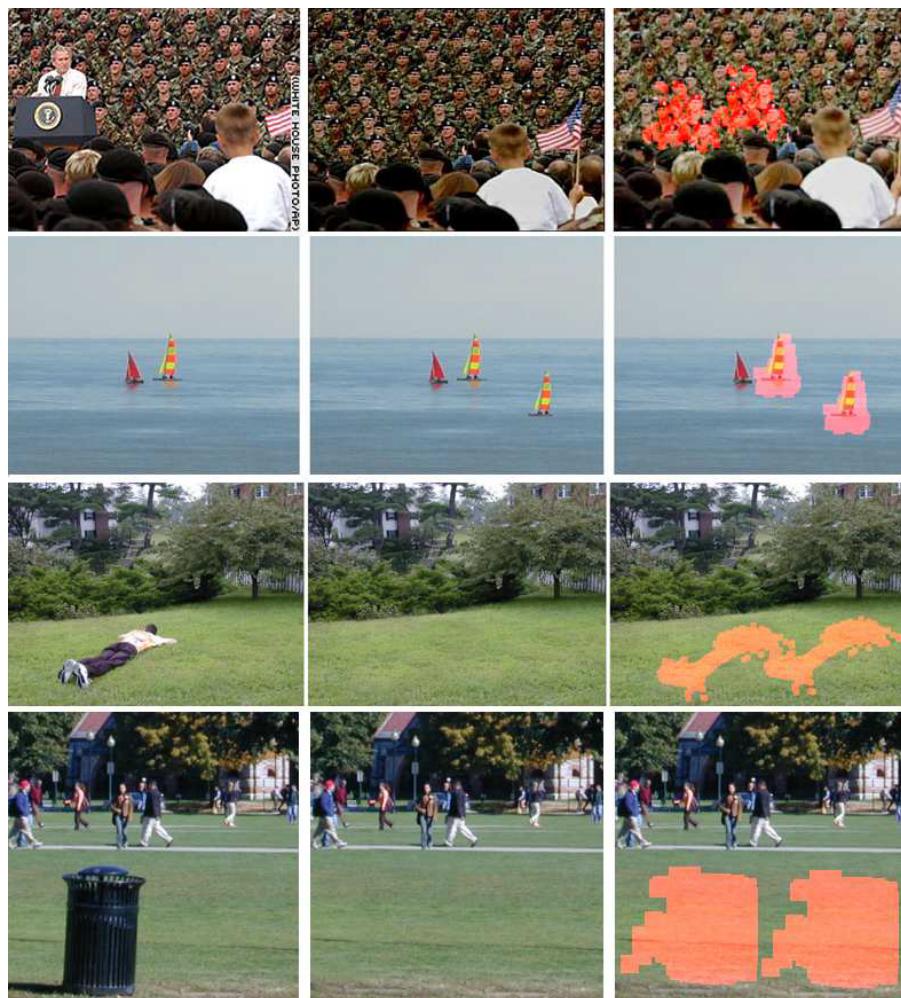


FIGURE 4.7 – Résultats de détection de duplications issues de Fig. 2 dans [5]. Chaque ligne présente une image originale, son altération par copie, puis la détection des régions dupliquées.

Ces méthodes classiques présentent plusieurs avantages : elles sont agnostiques au contenu visuel, ne nécessitent pas d'apprentissage, et permettent souvent une localisation précise des altérations. Toutefois, elles restent sensibles aux post-traitements (flou, compression, re-échantillonnage), et leur efficacité dépend fortement du choix des descripteurs utilisés. Avec l'évolution des outils de falsification, en particulier ceux fondés sur les GAN, la duplication peut être masquée de manière bien plus sophistiquée, ce qui appelle à renforcer ces approches par des méthodes plus robustes ou hybrides.

4.4 Limites face à l'augmentation de la complexité des falsifications

Pendant longtemps, les méthodes classiques de détection ont suffi à identifier la plupart des manipulations visuelles. Basées sur l'analyse d'artefacts, de duplications ou de ruptures statistiques, elles permettaient de repérer des modifications grossières ou mal intégrées. Pourtant ces approches montrent aujourd'hui leurs limites face à la sophistication croissante des falsifications plus récente.

Avec l'arrivée des modèles génératifs et en particulier des GAN, les possibilités de falsification ont radicalement évolué. Désormais, les images ou vidéos altérées ne se contentent plus d'éléments copiés-collés ou de retouches visible. Elles sont générées de toutes pièces, avec une cohérence visuelle qui dépasse les capacités des anciennes méthodes. Les artefacts classiques qui servaient de signaux d'alerte, bruit incohérent, contours flous et duplications suspectes peuvent aujourd'hui être volontairement simulés ou complètement effacés par les générateurs modernes.

Ce n'est donc plus seulement une question de détecter des erreurs, mais de faire face à des contenus synthétiques qui imitent avec précision les structures visuelles d'une scène authentique. Les techniques basées sur des heuristiques, aussi ingénieuses soient-elles, peinent à rivaliser avec des algorithmes capables d'apprendre les régularités du monde réel pour mieux les reproduire.

Dans ce contexte, un simple filtrage ou une analyse manuelle ne suffisent plus. Il devient indispensable de repenser les stratégies de détection, en intégrant des approches plus robustes, souvent basées sur l'apprentissage automatique. Ces nouvelles méthodes, capables d'extraire des indices bien plus subtils à partir de vastes ensembles de données annotées, constituent désormais une réponse plus adaptée à ces menaces.

Dans la suite de ce mémoire, nous nous tournerons donc vers ces techniques récentes,

en nous intéressant aux méthodes spécifiques conçues pour détecter les contenus générés par des GAN ainsi qu'aux défis qu'elles posent en pratique.

Chapitre 5

Modèles génératifs : principes et évolution avant les GAN

5.1 Modèles génératifs vs discriminatifs : définitions

On distingue généralement deux grandes familles de modèles statistiques, en fonction de leur finalité : les modèles modèle discriminatif, d'une part, et les modèles modèle génératif, d'autre part [25]. Cette classification bien que schématique permet de clarifier les grands paradigmes d'apprentissage et d'évaluer leur pertinence selon les types de tâches visées.

Les modèles discriminatifs ont pour objectif d'estimer la probabilité conditionnelle $P(y | x)$, c'est-à-dire la probabilité d'observer une sortie y sachant une entrée x . Ils sont principalement conçus pour des tâches de classification ou de prédiction, en apprenant à séparer les différentes catégories à partir des données observées. Des modèles tels que la régression logistique, les *Support Vector Machine* (SVM), ou les réseaux de neurones appliqués à des jeux de données étiquetés relèvent de cette approche. Leur efficacité provient du fait qu'ils se concentrent exclusivement sur la frontière de décision, sans chercher à modéliser la structure intrinsèque des données.

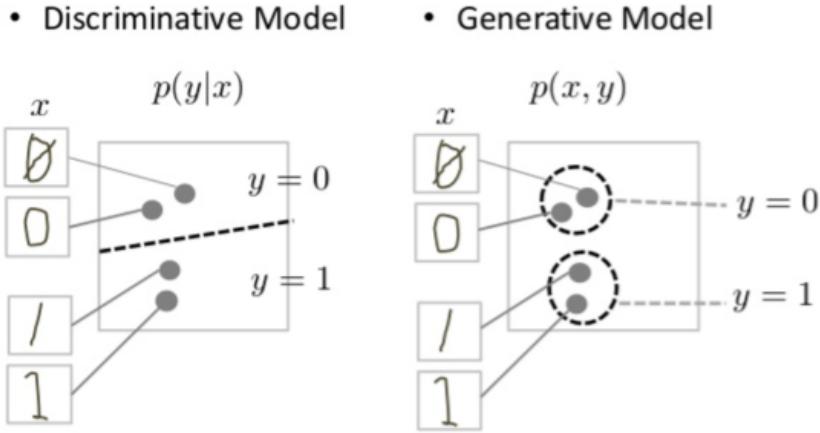


FIGURE 5.1 – Différences entre un modèle discriminatif ($P(y | x)$) et un modèle génératif ($P(x, y)$), dans une tâche de classification de chiffres manuscrits.

À l'inverse, les modèles génératifs cherchent à reproduire le processus qui a donné naissance aux données. Ils tentent d'approcher la distribution conjointe $P(x, y)$, ou, dans le cas non supervisé, la distribution marginale $P(x)$ [25]. Une telle modélisation permet non seulement de produire de nouveaux exemples réalistes, mais aussi de remplir des données manquantes, de réaliser des interpolations ou d'exécuter des tâches de denoising et de représentation non supervisée.

Ce type d'approche repose sur une hypothèse : les observations du monde réel sont considérées comme des échantillons tirés d'une distribution de probabilité sous-jacente $p(x)$. L'objectif du modèle est alors de construire une distribution paramétrée $q_\theta(x)$, dont les caractéristiques s'approchent de celles de $p(x)$, en minimisant une divergence statistique entre les deux. Une des formulations classiques repose sur la divergence de Kullback-Leibler $D_{\text{KL}}(p\|q_\theta)$ [25] :

$$\mathcal{L}(\theta) = \arg \min_{\theta} D(p\|q_\theta(x)) \quad (5.1)$$

Cette perspective probabiliste unifie plusieurs approches d'apprentissage non supervisé. En capturant la structure des données, les modèles génératifs sont capables de faire émerger des représentations latentes, de générer des exemples contre-factuels, ou encore d'apprendre des distributions conditionnelles complexes (par exemple $p(x | z)$ avec une variable latente z).

Toutes les approches de génération ne s'inscrivent pas dans ce cadre strictement probabiliste. Certaines méthodes comme le transfert de style ou la synthèse de texture [6], s'appuient davantage sur des objectifs perceptifs définis par des réseaux pré-entraînés. D'autres modèles, comme les réseaux adversariaux créatifs [33], adoptent

volontairement une stratégie de déviation par rapport à la distribution d'origine, en générant des contenus qui sortent du domaine des données réelles.

Bien avant l'émergence des GAN, plusieurs familles de modèles probabilistes ont été développées pour approximer la distribution des données : auto-encodeurs (auto-encodeur (AE)), variantes variationnelles (auto-encodeur variationnel (VAE)), modèles à base de chaînes de Markov (RBM, *Hidden Markov Model* (HMM)), etc. La section suivante reviendra en détail sur ces approches « classiques », qui sont les bases théoriques et pratiques de la génération d'images avant les architectures adversariales.

5.2 Modèles classiques : auto-encodeurs, VAE, modèles de Markov

Avant l'apparition des GAN, plusieurs familles de modèles génératifs ont été proposées pour approximer la distribution d'une variable aléatoire complexe. Ces modèles sont fondés sur des principes probabilistes ou sur des techniques de reconstruction. Cette section revient sur les trois catégories principales : les auto-encodeurs standards, les auto-encodeurs variationnels et les modèles de Markov.

5.2.1 Auto-encodeurs (AE)

Les AE sont des architectures de réseaux de neurones non probabilistes conçues pour apprendre une représentation compressée d'une donnée. Un auto-encodeur se compose de deux parties : un encodeur f_θ qui projette l'entrée x dans un espace latent z , et un décodeur g_ϕ qui reconstruit une estimation x' à partir de ce vecteur latent :

$$z = f_\theta(x), \quad x' = g_\phi(z)$$

où g_ϕ est le décodeur qui reconstruit une estimation $x' \in \mathbb{R}^n$ de l'entrée initiale à partir de cette représentation. Dans le cadre d'un auto-encodeur classique, x et x' sont de même dimension, ce qui permet d'évaluer l'erreur de reconstruction.

L'apprentissage se fait par la minimisation d'une fonction de perte de reconstruction, typiquement la distance quadratique moyenne entre x et x' :

$$\mathcal{L}(\theta, \phi) = \mathbb{E}_{x \sim p_{\text{data}}} [\|x - g_\phi(f_\theta(x))\|^2]$$

où $\mathbb{E}_{x \sim p_{\text{data}}}$ désigne l'espérance calculée sur la distribution réelle des données p_{data} . En pratique, cette espérance est approchée par une moyenne empirique sur les

exemples du jeu d’entraînement.

Même si les auto-encodeurs peuvent être utilisés pour générer de nouvelles données en échantillonnant des points dans l’espace latent, ils ne sont pas conçus pour le faire de manière probabiliste. L’espace latent n’est généralement pas structuré, ce qui rend l’interpolation ou la génération difficile à contrôler.

5.2.2 Auto-encodeurs variationnels (VAE)

Les VAE, proposés comme une extension probabiliste des AE, introduisent une modélisation explicite de l’espace latent à l’aide d’une variable aléatoire z , régie par une loi a priori $p(z)$, souvent choisie comme une distribution normale centrée réduite $\mathcal{N}(0, I)$. Contrairement aux AE classiques, qui encodent chaque donnée x en un point unique dans l’espace latent, les VAE associent à chaque entrée une distribution sur z , ce qui permet une génération plus souple et probabiliste.

Dans ce cadre, l’encodeur n’est plus une simple projection, mais une approximation de la distribution postérieure $q_\phi(z | x)$, généralement paramétrée par un réseau de neurones produisant une moyenne et une variance pour chaque z . Le décodeur, quant à lui, modélise la vraisemblance $p_\theta(x | z)$, c’est-à-dire la probabilité de reconstruire une donnée x à partir d’un vecteur latent z .

L’apprentissage repose sur la maximisation d’une borne inférieure du logarithme de la vraisemblance marginale (ELBO), qui constitue l’objectif optimisé durant l’entraînement. Cette borne inférieure est notée $\mathcal{L}(\theta, \phi)$, et elle est construite de manière à approcher au mieux la log-vraisemblance réelle des données. Maximiser l’ELBO revient ainsi à rendre le modèle génératif aussi proche que possible de la distribution des données observées :

$$\mathcal{L}(\theta, \phi) = \mathbb{E}_{q_\phi(z|x)} [\log p_\theta(x | z)] - D_{\text{KL}}(q_\phi(z | x) \| p(z))$$

Le premier terme mesure la qualité de reconstruction des données, tandis que le second — la divergence de Kullback-Leibler — agit comme une régularisation en forçant la distribution $q_\phi(z | x)$ à rester proche du prior $p(z)$. Pour maximiser cette fonction, on cherche donc à obtenir des reconstructions précises (terme de log-vraisemblance élevé) tout en minimisant la divergence avec la distribution latente cible. Cette contrainte impose une structure régulière à l’espace latent \mathcal{Z} , ce qui permet notamment d’en échantillonner directement des vecteurs $z \sim p(z)$ et de générer de nouvelles données à l’aide du décodeur.

Cette organisation rend également possibles des opérations comme l’interpolation dans l’espace latent, c’est-à-dire la génération de données intermédiaires entre deux exemples en moyennant leurs vecteurs latents, de manière fluide et contrôlée.

On parle dans ce contexte de « vraisemblance amortie », car la distribution postérieure approximative $q_\phi(z | x)$ est inférée par un seul réseau de neurones partagé pour l’ensemble des données. Cela rend l’inférence rapide et compatible avec des jeux de données volumineux, mais limite aussi la capacité du modèle à s’adapter finement à chaque exemple. En conséquence, les VAE peuvent produire des échantillons visuellement plus flous, en partie à cause de cette approximation globale, ainsi que de la dispersion induite par l’échantillonnage [25].

5.2.3 Modèles de Markov et chaînes latentes

Les modèles de Markov constituent une autre famille historique de modèles génératifs, reposant sur la factorisation de la distribution jointe d’une séquence en une succession de probabilités conditionnelles. Dans les cas les plus simples, comme les HMM, on suppose l’existence d’une chaîne de variables cachées z_t régies par une dynamique de Markov :

$$P(z_t | z_{t-1}) = P(z_t | z_{t-1}, z_{t-2}, \dots) \quad (\text{propriété de Markov})$$

Les observations x_t sont alors conditionnées sur les états latents z_t via une distribution d’émission $P(x_t | z_t)$. Ces modèles sont particulièrement adaptés aux données séquentielles (texte, audio, vidéo), et permettent à la fois de modéliser et de générer des séquences réalistes, mais leur expressivité reste limitée aux dynamiques locales de faible ordre, sauf à introduire des structures plus complexes ou des approximations variationnelles.

Les modèles de type RBM (Restricted Boltzmann Machines) et leurs extensions, bien que plus puissants, souffrent également d’une difficulté d’entraînement liée à la nature non tractable de la distribution jointe modélisée. L’apprentissage repose souvent sur des approximations comme le contraste divergence. Ces limitations ont ouvert la voie à des modèles plus flexibles, comme les GAN, capables de générer des données de qualité visuelle supérieure sans modéliser explicitement une fonction de vraisemblance.

5.3 Premières tentatives de génération d’images

Avant l’essor des GAN, plusieurs approches ont tenté de s’attaquer à la génération automatique d’images, avec des objectifs, des structures et des résultats très variés. Ces premiers modèles se fondaient soit sur des principes probabilistes (comme les VAE ou les modèles de Markov), soit sur des heuristiques perceptives ou des stratégies de reconstruction, sans modélisation explicite de la vraisemblance des données.

Les auto-encodeurs standards, bien qu’orientés principalement vers la compression et la reconstruction, ont été rapidement détournés à des fins de génération. En interpolant entre les représentations latentes de différentes images, il devenait possible de produire des formes visuelles nouvelles. Toutefois, l’absence de contrôle probabiliste sur l’espace latent rendait ces interpolations imprévisibles et souvent peu réalistes.

Les VAE ont permis un pas important en direction de la génération contrôlée. Grâce à leur formulation probabiliste et à la régularisation imposée sur l’espace latent, ils permettent d’échantillonner directement des vecteurs latents à partir d’un prior connu, puis de générer des images plausibles via le décodeur. Néanmoins comme évoqué précédemment, la qualité visuelle reste limitée CAR souvent caractérisée par des contours flous ou des textures imprécises [25].

D’autres approches plus heuristiques ont tenté de produire des images à partir de contraintes stylistiques ou de représentations visuelles haut niveau. L’algorithme de transfert de style neuronal proposé par Gatys et al. [6] a démontré qu’il était possible de recomposer une image en imitant le style d’une autre, en exploitant les couches intermédiaires d’un réseau convolutif pré-entraîné. Ce type de méthode ne repose pas sur une distribution générative au sens strict mais sur une ré-optimisation d’une image en fonction d’un critère de similarité perceptuelle. Ce qui n’est pas génératif au sens probabiliste du terme.

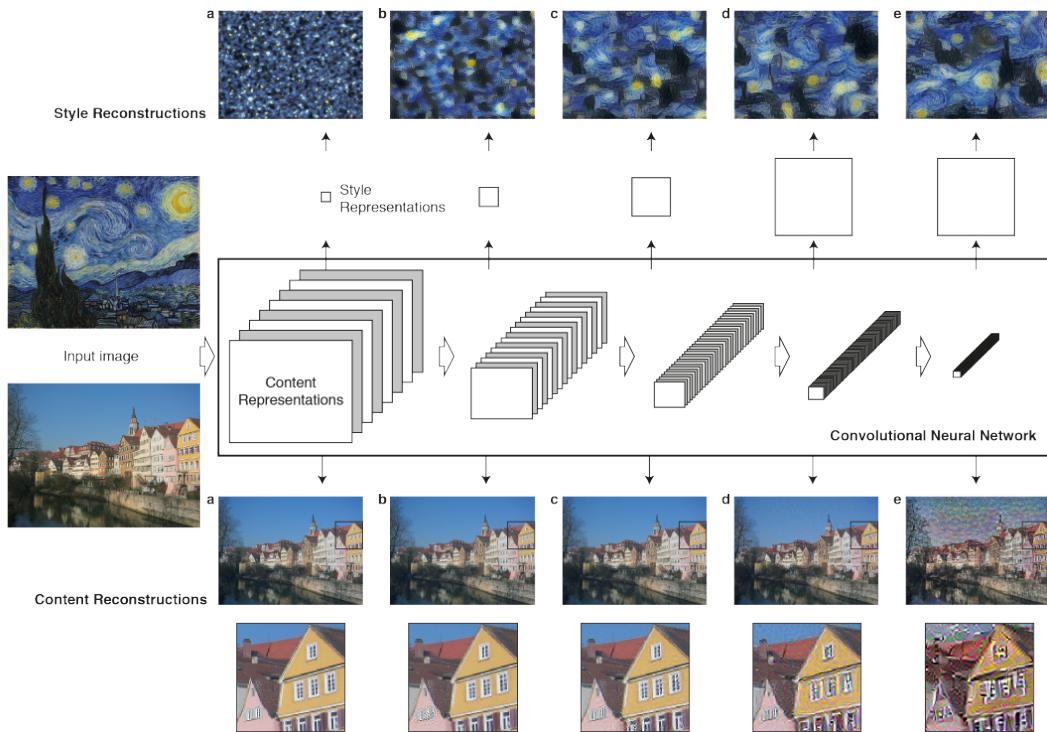


FIGURE 5.2 – Schéma du processus de transfert de style neuronal. L'image d'entrée (en bas à gauche) est traitée par un réseau convolutionnel pour en extraire les représentations de contenu, tandis que l'image de style (en haut à gauche) fournit des représentations stylistiques. Les couches du réseau sont exploitées pour recomposer une nouvelle image combinant contenu et style à différents niveaux d'abstraction [6].

Enfin, les modèles d'énergie tels que les RBM ou les Deep Belief Networks ont également été explorés pour la génération, notamment d'images binaires ou de petites dimensions. Comme illustré par les échantillons générés à partir de MNIST, ces modèles produisent des résultats grossiers et peu définis visuellement. Malgré quelques succès conceptuels, les difficultés d'entraînement et le coût élevé de l'échantillonnage par chaînes de Markov ont rapidement montré leurs limites pour des tâches de génération réaliste à grande échelle.

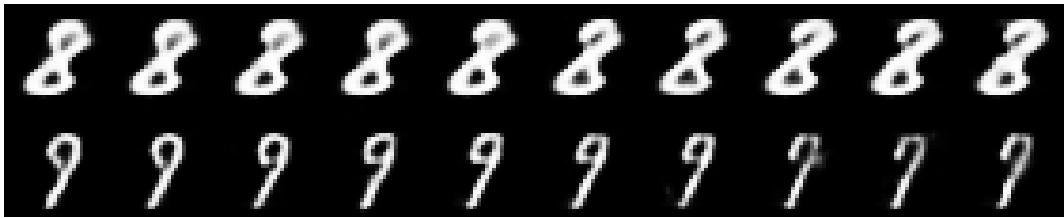


FIGURE 5.3 – Séquences d’images générées par un RBM entraîné sur MNIST. Entre chaque paire d’images, une étape de Gibbs sampling est effectuée ($v \rightarrow h \rightarrow v$). La première ligne illustre une transformation progressive d’un chiffre 8, la seconde une transition du chiffre 9 vers 7. [7].

Ces premières tentatives ont posé des bases importantes tant théoriques que méthodologiques. Elles ont permis de mieux comprendre les exigences propres à la génération d’images réalistes qui sont le contrôle de l’espace latent, stabilité de l’apprentissage, fidélité des détails visuels. Ces limitations seront précisément celle que les GAN chercheront à dépasser avec l’apprentissage adversarial.

5.4 Limites des approches pré-GAN

Malgré leur diversité, les méthodes de génération antérieures aux GAN présentent plusieurs limitations majeures, tant du point de vue théorique que pratique. Ces freins ont justifié l’émergence de nouveaux paradigmes plus adaptés à la synthèse d’images réalistes à grande échelle.

5.4.1 Expressivité restreinte des modèles probabilistes explicites

Les modèles tels que les VAE, les modèles autoregressifs ou encore les modèles de Markov tentent de modéliser directement la distribution des données $p(x)$. Cela suppose souvent des hypothèses fortes sur la structure ou la factorisation de cette distribution, limitant leur capacité à capturer des phénomènes visuels complexes. Les VAE, par exemple, utilisent une distribution latente souvent gaussienne isotrope, ce qui contraint la diversité des échantillons générés [25]. De même, les modèles autoregressifs comme PixelRNN ou PixelCNN génèrent les pixels un par un, ce qui entraîne une lourdeur computationnelle et une dépendance temporelle parfois contre-productive.[34]

5.4.2 Qualité visuelle insuffisante

Un reproche fréquent adressé aux approches probabilistes classiques concerne la qualité visuelle des images générées. Les VAE malgré leur efficacité pour échantillonner des représentations latentes structurées, produisent souvent des images floues. Ce flou est lié à l’objectif d’optimisation de la vraisemblance qui favorise des moyennes

d’images plausibles plutôt que des détails nets. Quant aux modèles d’énergie comme les RBMs, leurs capacités de génération se confrontent rapidement à des limites dès que la complexité visuelle augmente. Les images produites à partir de MNIST en sont un exemple : malgré une certaine cohérence elles restent grossières et manquent de fidélité visuelle [7].

5.4.3 Difficultés d’entraînement et de convergence

Les modèles d’énergie en particulier les RBMs et les Deep Belief Network nécessitent des techniques d’entraînement spécialisées comme l’échantillonnage par Gibbs et l’algorithme de contraste divergence. Ces méthodes introduisent un coût computationnel élevé et des problèmes de convergence surtout pour des données en haute dimension [25]. Même les VAE peuvent souffrir de phénomènes comme le *posterior collapse*, où le réseau encodeur ignore la variable latente, rendant la génération peu informative.

5.4.4 Absence de critère perceptuel ou discriminatif

Enfin, la plupart de ces approches cherchent uniquement à minimiser une divergence statistique entre la distribution modèle et celle des données réelles (comme la divergence de Kullback-Leibler), sans considérer la qualité perçue par un humain. Aucune contrainte ne pousse directement le modèle à générer des images qui soient crédibles pour un humain. Les GAN à l’inverse introduiront une fonction de coût adversariale qui repose sur la discrimination entre le vrai et le faux, offrant ainsi une régularisation perceptive implicite.

L’expressivité restreinte, la qualité visuelle souvent décevante, le coût d’entraînement important et l’absence de prise en compte de critères perceptuels ont progressivement mis en évidence les limites des approches traditionnelles. Ces constats ont ouvert la voie à de nouvelles architectures comme les GAN qui seront pensés pour dépasser ces limites conceptuelles et techniques.

Chapitre 6

Les GAN : fondements, évolutions et capacités de génération

6.1 Principes fondamentaux des GAN

6.1.1 Principe du jeu adversarial

Les GAN, introduits en 2014 par Ian Goodfellow et al. [8], ont profondément transformé l'approche de la génération de données synthétiques par apprentissage automatique. Leur originalité réside dans l'organisation d'un jeu à somme nulle opposant deux réseaux de neurones : un générateur G qui produit des données artificielles, et un discriminateur D chargé de distinguer les échantillons synthétiques des véritables données issues du jeu d'entraînement.

Le générateur apprend à transformer un vecteur latent aléatoire \mathbf{z} , issu d'une distribution simple (comme une gaussienne multivariée), en une image réaliste. Le discriminateur, quant à lui, reçoit une image (réelle ou générée) et prédit la probabilité qu'elle provienne des données réelles. Ce jeu compétitif pousse chaque réseau à s'améliorer mutuellement, le générateur en rendant ses images plus convaincantes, le discriminateur en affinant sa capacité de détection.

6.1.2 Formulation mathématique

La dynamique entre le générateur G et le discriminateur D se formalise comme un problème d'optimisation à deux joueurs, dans le cadre d'un jeu minimax. Le discriminateur cherche à distinguer les données réelles des données générées en maximisant une fonction objectif, tandis que le générateur tente de minimiser cette même fonction, en produisant des échantillons suffisamment convaincants pour tromper le discriminateur.

Cette interaction est décrite par la fonction d'objectif, que l'on cherche à optimiser :

$$\min_G \max_D V(D, G) = \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}} [\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{z} \sim p_{\mathbf{z}}} [\log(1 - D(G(\mathbf{z})))] , \quad (6.1)$$

où p_{data} désigne la distribution réelle des données d'apprentissage, $p_{\mathbf{z}}$ désigne la distribution dans l'espace latent, dont la nature a été discutée dans le chapitre précédent sur les modèles génératifs, $G(\mathbf{z})$ représente un échantillon synthétique généré à partir du bruit latent \mathbf{z} , et $D(x)$ donne une estimation de la probabilité que l'entrée x provienne de la distribution réelle, autrement dit $D(x) \approx P(y = 1 | x)$.

La fonction $V(D, G)$ repose sur deux composantes : le premier terme, $\mathbb{E}_{\mathbf{x} \sim p_{\text{data}}} [\log D(\mathbf{x})]$, pousse le discriminateur à attribuer une probabilité élevée aux vraies données. Le second, $\mathbb{E}_{\mathbf{z} \sim p_{\mathbf{z}}} [\log(1 - D(G(\mathbf{z})))]$, l'incite à rejeter les échantillons produits par le générateur.

On peut formaliser ce cadre en introduisant une variable binaire latente $y \in \{0, 1\}$, avec $y = 1$ si l'échantillon x est réel ($x \sim p_{\text{data}}$), et $y = 0$ s'il est synthétique, c'est-à-dire généré par G . L'apprentissage du discriminateur peut alors être vu comme une tâche de classification supervisée, dont les labels sont construits artificiellement : toutes les données réelles sont étiquetées positives (classe 1), tandis que les données générées sont étiquetées négatives (classe 0).

De son côté, le générateur cherche à produire des images telles que $D(G(\mathbf{z}))$ soit aussi proche que possible de 1, autrement dit que ses productions soient classées comme vraies par le discriminateur. Ce faisant, il tente d'induire une erreur de classification : en minimisant $V(D, G)$, il maximise implicitement la confusion du discriminateur, et donc sa propre capacité à produire des données crédibles.

Ce qui donne lieu à un jeu compétitif dans lequel les deux réseaux sont en tension permanente : le discriminateur apprend à mieux détecter les faux, tandis que le générateur affine ses productions pour échapper à la détection. Ce mécanisme

constitue le cœur de l'apprentissage adversarial.

Dans le cas idéal, capacité infinie des deux réseaux, optimisation parfaite, et convergence assurée, le générateur parvient à reproduire exactement la distribution des données réelles, c'est-à-dire $p_g = p_{\text{data}}$. *Le discriminateur, alors privé d'informations exploitables, est contraint de rendre une prédiction constante pour toute entrée.* [8] Goodfellow et al. [8] démontrent que cet équilibre correspond à l'optimum global du jeu, atteint lorsque $p_g = p_{\text{data}}$. Le discriminateur optimal dans ce cas est donné par :

$$D_G^*(x) = \frac{p_{\text{data}}(x)}{p_{\text{data}}(x) + p_g(x)}. \quad (6.2)$$

Lorsque les deux distributions coïncident parfaitement, on obtient alors :

$$\forall x, D_G^*(x) = 0,5 \quad (6.3)$$

Ce qui formalise l'indistinguabilité recherchée entre vraies et fausses données.

Ce point d'équilibre constitue le but théorique du jeu : une situation où le générateur a parfaitement appris la distribution réelle, rendant ses échantillons indistinguables des vraies données. Ce point correspond en théorie des jeux, à un équilibre de Nash : aucun des deux joueurs, ni le générateur et ni le discriminateur, ne peut améliorer son gain en modifiant sa stratégie de manière unilatérale [8]. Le jeu atteint ainsi un état stable où les deux réseaux ont atteint leur performance optimale par rapport à leur adversaire.

Enfin, il est important de noter que cette formulation équivaut, d'un point de vue théorique, à la minimisation de la divergence de JSD (Jensen–Shannon divergence) entre la distribution réelle et celle générée. Cette mesure symétrique atteint son minimum lorsque $p_g = p_{\text{data}}$, ce qui justifie mathématiquement l'objectif poursuivi par le processus d'apprentissage.

Ce cadre mathématique, bien que conceptuellement élégant, soulève en pratique de nombreuses difficultés d'optimisation, notamment liées à la stabilité du jeu entre les deux réseaux. Ces enjeux seront abordés dans la sous-section suivante.

6.1.3 Défis pratiques de l'entraînement

Malgré son élégance théorique, entraîner un GAN de manière stable demeure difficile. En pratique, l'équilibre entre le générateur et le discriminateur est délicat à maintenir. Lorsque le générateur est encore peu performant, ses échantillons sont aisément détectés, ce qui conduit à une quasi-certitude de classification par le discriminateur. Cela provoque une saturation du gradient et un affaiblissement du

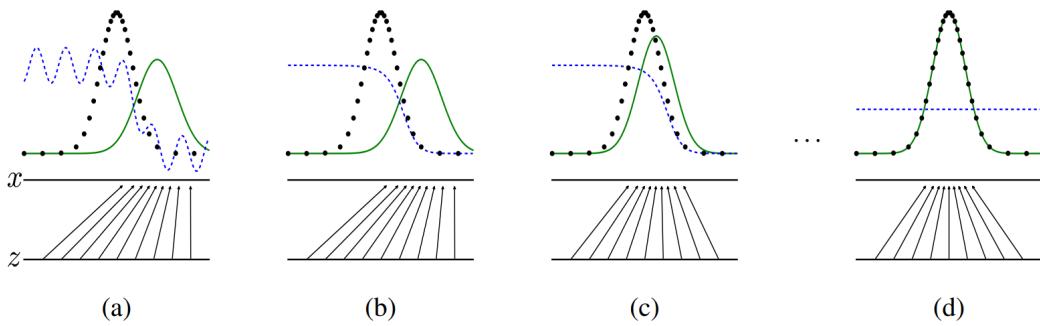


FIGURE 6.1 – Illustration du mécanisme d'apprentissage des GANs : au fil de l'entraînement, le générateur améliore ses productions tandis que le discriminateur affine sa frontière de décision, jusqu'à ce que les distributions synthétique et réelle deviennent indiscernables. [8]

signal d'apprentissage. À l'inverse, si le discriminateur devient trop efficace trop tôt, il cesse de fournir un retour utile au générateur, bloquant sa progression.

Pour atténuer ces effets, plusieurs ajustements ont été proposés dès les premiers travaux. Une stratégie classique consiste à modifier la fonction objectif initiale du générateur. Plutôt que de minimiser $\log(1 - D(G(z)))$, une expression dont la dérivée devient négligeable lorsque $D(G(z)) \rightarrow 0$, on préfère maximiser directement $\log D(G(z))$, ce qui fournit un gradient plus informatif, notamment en début d'apprentissage. Cette reformulation, dite « non saturante », peut être exprimée comme suit :

$$\begin{aligned} J_{\text{minimax}}^{(G)} &= \mathbb{E}_{z \sim p_z} [\log (1 - D(G(z)))] \\ J_{\text{NS}}^{(G)} &= -\mathbb{E}_{z \sim p_z} [\log D(G(z))] \end{aligned} \quad (6.4)$$

Bien que ces deux formulations mènent théoriquement au même optimum, seule la deuxième évite les problèmes de saturation des gradients, ce qui la rend plus adaptée en pratique.

Ces ajustements sont d'autant plus cruciaux que l'entraînement des GANs repose sur une optimisation alternée par descente de gradient stochastique sur minibatch. Ce protocole, décrit dans l'algorithme 1 de l'article fondateur de Goodfellow et al. [8], consiste à effectuer plusieurs mises à jour du discriminateur pour chaque mise à jour du générateur. Ce déséquilibre contrôlé permet au discriminateur de mieux guider l'apprentissage de G , mais il peut aussi agraver les instabilités si les fréquences d'actualisation ou les taux d'apprentissage sont mal choisis.

D'autres techniques ont été explorées pour stabiliser la dynamique du jeu : introduction de bruit sur les entrées du discriminateur pour limiter l'overfitting, équilibrage

fin du rythme d'optimisation entre D et G , ou encore recours à des régularisations explicites comme les gradient penalties. La diversité de ces méthodes reflète la complexité intrinsèque du processus adversarial.

Parmi les difficultés récurrentes, le mode collapse est l'un des plus emblématiques. Dans ce cas, le générateur apprend à produire un ensemble restreint d'échantillons (voire une seule image) qui trompe efficacement le discriminateur, mais au détriment de la diversité. Ce phénomène révèle l'existence d'un équilibre local sous-optimal, où le générateur exploite une faiblesse momentanée de son adversaire sans parvenir à généraliser à toute la distribution réelle.

Les travaux de Fedus et al. [35] ont remis en question l'idée selon laquelle l'apprentissage des GANs consisterait à minimiser explicitement une divergence (comme la divergence de Jensen–Shannon) à chaque étape. Les auteurs montrent empiriquement que même dans des situations où cette divergence ne fournit plus de gradient exploitable, certaines formulations alternatives, comme la loss non saturante, continuent de guider efficacement l'apprentissage. Cela s'explique notamment par leur capacité à amplifier les différences de sortie du discriminateur, même lorsqu'elles sont infimes, fournissant ainsi un signal d'apprentissage robuste.

Ces constats ont favorisé l'émergence de nombreuses variantes architecturales ou fonctionnelles de GANs. Nous présenterons dans la section suivante certaines de ces évolutions pertinentes dans le cadre de la détection de faux contenus.

6.2 Architecture de base et variantes majeures

Bien que tous les GAN reposent sur un même principe d'apprentissage adversarial entre un générateur et un discriminateur, les performances concrètes des modèles dépendent fortement de leur architecture. Depuis la proposition initiale formulée par Ian Goodfellow et al. [8], de nombreuses variantes ont vu le jour afin d'améliorer la stabilité de l'entraînement, la qualité visuelle des images générées, ou encore la capacité à représenter des distributions complexes de manière plus fidèle.

Dans cette section, nous décrivons trois architectures particulièrement influentes qui ont marqué l'évolution des GANs dans le domaine de la synthèse d'image. Les *Deep Convolutional Generative Adversarial Network* (DCGAN) qui instaurent des bonnes pratiques architecturales pour l'usage de convolutions profondes, les *Progressive Growing of GANs* (ProGAN), qui introduisent une montée en résolution progressive et les *Style-based Generator Architecture for GANs* (StyleGAN) qui offrent un contrôle explicite sur l'apparence des images produites. Chacune de ces avancées a contribué

à franchir un seuil qualitatif dans la génération d'images et donc par conséquence de nouveaux défis pour la détection de ceux-ci.

6.2.1 DCGAN : vers une architecture convolutive stable

Les premiers modèles de GAN, bien que prometteurs, souffraient d'une instabilité notable à l'entraînement, en particulier lorsque les réseaux profitait d'une certaine profondeur. C'est dans ce contexte que Radford et al. [9] ont proposé en 2016 une architecture spécifique : DCGAN, destinée à stabiliser l'apprentissage tout en tirant parti des propriétés hiérarchiques des réseaux convolutifs profonds. Leur travail a posé les bases d'une nouvelle classe de GANs plus robustes et plus interprétables, en imposant une série de contraintes architecturales simples mais efficaces.

Principes architecturaux

L'originalité des DCGAN réside dans la combinaison de techniques issues des CNN classiques avec les exigences propres à l'apprentissage adversarial. Les auteurs identifient plusieurs choix structurants pour stabiliser la dynamique entre G et D :

- remplacement des couches de pooling par des convolutions à pas variable (strided ou fractionally-strided convolutions),
- suppression des fully connected layers au profit d'un empilement convolutionnel pur,
- normalisation par lot (*batch normalization*) dans presque toutes les couches, à l'exception de l'entrée de D et de la sortie de G ,
- utilisation de ReLU dans G (sauf en sortie, où l'on emploie *tanh*), et de LeakyReLU dans D ,

Ces principes permettent non seulement une meilleure stabilité de l'entraînement, mais favorisent également l'apprentissage de représentations intermédiaires exploitables. Le générateur est ainsi capable de produire des images plus cohérentes, tandis que le discriminateur apprend des caractéristiques visuelles interprétables, même en l'absence de supervision explicite.

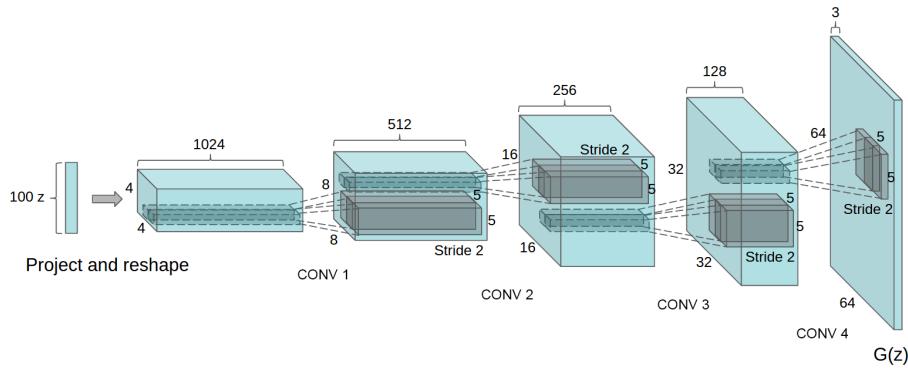


FIGURE 6.2 – Architecture du générateur DCGAN utilisé pour modéliser les scènes du jeu de données LSUN. Le vecteur latent est projeté, puis transformé par une série de convolutions transposées en une image 64×64 sans fully connected layer [9].

Représentations apprises et applications

Un apport majeur des DCGAN est d'avoir démontré que les représentations internes apprises, notamment dans le discriminateur, sont réutilisables pour d'autres tâches. Cette capacité à capturer des caractéristiques visuelles pertinentes, même sans supervision explicite, marque une avancée notable dans l'utilisation des GANs comme outils de représentation.

Feature maps discriminatives En réutilisant les feature maps issues des couches convolutives de D comme descripteurs visuels, les auteurs montrent que ces représentations permettent d'atteindre des performances compétitives sur des tâches de classification supervisée, telles que CIFAR-10 ou SVHN, en n'utilisant qu'un classifieur linéaire en aval.

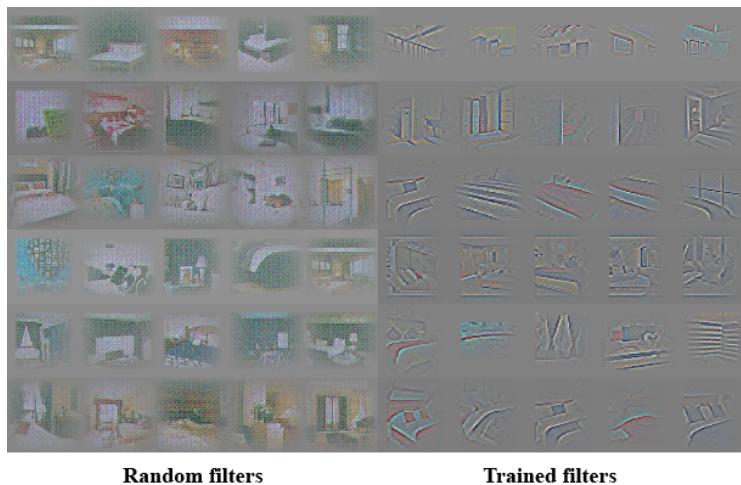


FIGURE 6.3 – Visualisation guidée des activations de couches du discriminateur entraîné sur LSUN. Certaines activations sont spécifiques à des éléments visuels (lits, fenêtres) [9].

Structure de l'espace latent Les DCGAN révèlent également une organisation sémantique intéressante dans l'espace latent Z . Radford et al. [9] mettent en lumière des interpolations continues entre vecteurs latents, traduisant des transformations progressives dans les images générées.

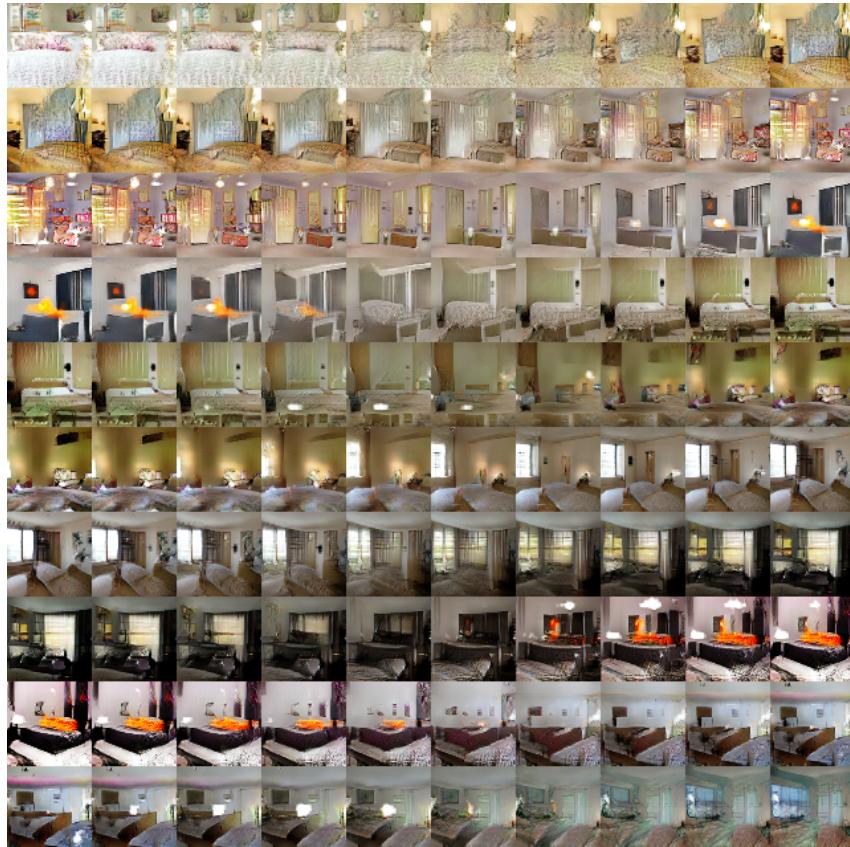


FIGURE 6.4 – Interpolation entre vecteurs latents dans l'espace Z : les transitions entre chambres générées sont progressives et cohérentes [9].

Il est également possible d'effectuer des opérations vectorielles simples pour manipuler des attributs visuels spécifiques. Par exemple, en ajoutant ou soustrayant certains vecteurs latents, on peut induire des modifications telles que l'apparition de lunettes, un sourire ou l'orientation du visage. Cela suggère une structuration hiérarchique et sémantique de l'espace latent, apprise sans supervision.

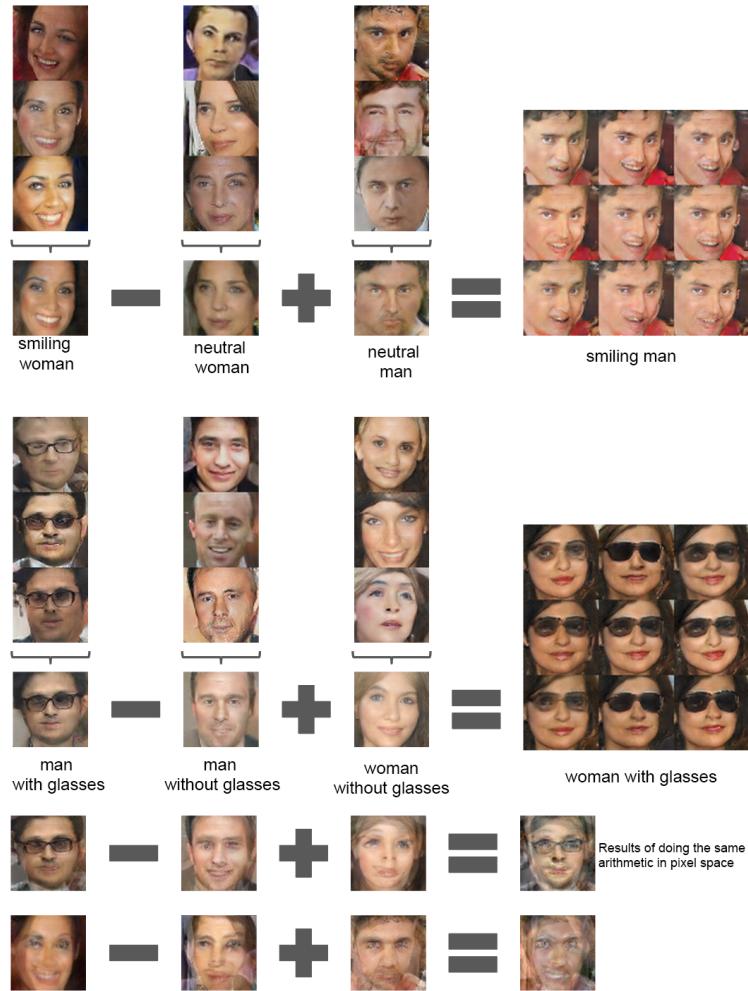


FIGURE 6.5 – Exemple d’arithmétique vectorielle dans l’espace latent des DCGAN. Des opérations linéaires sur les vecteurs latents permettent de manipuler des attributs visuels spécifiques [9].

Limites et apports

L’introduction des DCGAN a représenté une avancée importante dans la conception de GANs plus stables et exploitables. Grâce à leur architecture convolutionnelle dépourvue de fully connected layers, ces modèles ont permis d’améliorer la qualité visuelle des échantillons générés tout en facilitant l’entraînement de réseaux plus profonds. Les images produites, même sur des jeux de données complexes, témoignent d’une meilleure cohérence globale et d’un meilleur réalisme par rapport aux premiers GANs.

L’apprentissage de représentations intermédiaires utiles, visibles dans les activations du discriminateur et dans la structure de l’espace latent, a contribué à renforcer l’interprétabilité de ces modèles. La possibilité de réutiliser les feature maps comme

descripteurs visuels ou d’exploiter la linéarité de l’espace Z pour manipuler des attributs a constitué une nouveauté marquante dans la littérature.

Malgré leur succès initial, les DCGAN présentent plusieurs limites. Leur stabilité reste conditionnée à un réglage fin des hyperparamètres, et leur efficacité diminue nettement pour des résolutions plus élevées. Le réseau générateur peine alors à capturer la richesse des textures et à maintenir la cohérence spatiale sur des images de grande taille. Par ailleurs, les DCGAN demeurent vulnérables aux problèmes classiques des GANs, notamment le mode collapse, qui se manifeste par une faible diversité des échantillons générés.

L’architecture des DCGAN, bien que plus robuste, reste relativement rigide. Elle ne permet pas un contrôle fin sur l’apparence des images générées, contrairement aux modèles ultérieurs comme StyleGAN. De ce fait, les DCGAN constituent davantage une étape charnière qu’une solution définitive, marquant la transition vers des architectures plus sophistiquées et expressives.

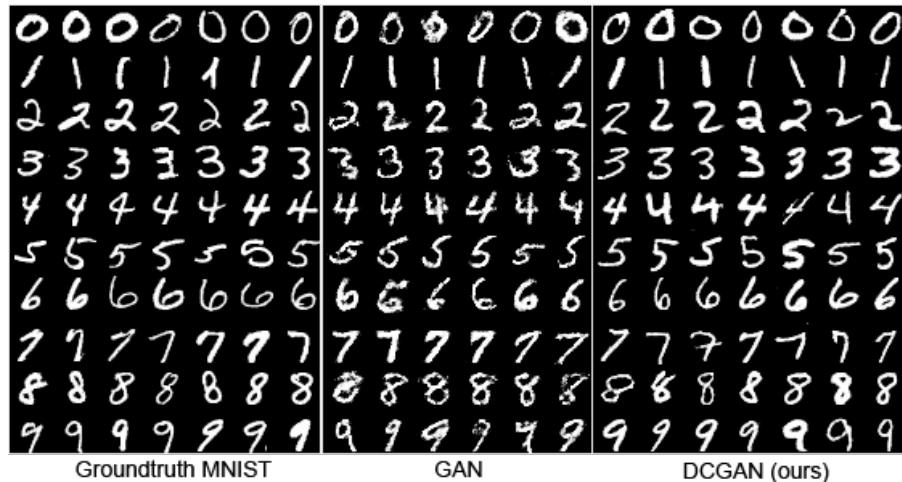


FIGURE 6.6 – Comparaison entre des échantillons générés par un GAN classique (milieu) et par un DCGAN • (droite), sur le dataset MNIST. Les images réelles sont affichées à gauche. [9]

6.2.2 ProGAN : génération progressive pour stabiliser la montée en résolution

Alors que les DCGAN avaient établi des bases architecturales stables pour la génération d’images, ils demeuraient limités à des résolutions modestes (64×64 ou 128×128 pixels), peinant à modéliser des structures complexes à plus haute échelle. C’est dans cette optique que Tero Karras et al. [10] ont proposé ProGAN, une toute nouvelle architecture qui a pour principe de faire croître progressivement la taille du générateur et du discriminateur pendant l’entraînement.

Principe de la croissance progressive

Cette approche repose sur un apprentissage progressif, par étapes. On commence par générer de petites images (par exemple 4×4), que l'on affine ensuite progressivement en ajoutant des couches qui augmentent la résolution : 8×8 , puis 16×16 , jusqu'à atteindre le plein format (par exemple 1024×1024 pour CelebA-HQ). À chaque étape, le réseau hérite des poids déjà appris, ce qui évite de repartir de zéro à chaque résolution.

Cette montée en complexité permet au générateur de se concentrer d'abord sur les grandes structures globales (formes du visage, position des yeux), puis d'ajouter progressivement des détails plus fins (texture de peau, cheveux, etc.). Une phase de transition avec interpolation linéaire entre anciennes et nouvelles couches assure une continuité fluide dans l'apprentissage.

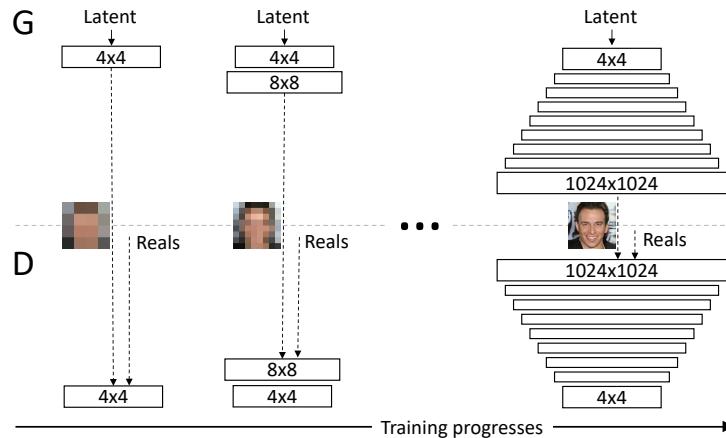


FIGURE 6.7 – Principe de la croissance progressive dans ProGAN : l'entraînement commence avec une faible résolution (4×4) à la fois pour le générateur (G) et le discriminateur (D). Des couches supplémentaires sont ensuite ajoutées progressivement, doublant à chaque étape la résolution spatiale des images synthétisées. Ce processus favorise une génération stable à haute résolution [10].

Bénéfices et apports techniques

Outre la stabilisation de l'entraînement en haute résolution, cette stratégie apporte plusieurs avantages :

- une réduction du temps d'entraînement : la majorité des itérations étant effectuée sur de faibles résolutions, les calculs sont moins coûteux,
- une amélioration de la convergence : chaque bloc convolutif apprend sur une tâche de complexité croissante, facilitant l'optimisation,
- une meilleure qualité visuelle : les images finales sont plus nettes, plus cohérentes, avec une plus grande diversité.

Tero Karras et al. [10] introduisent également deux techniques complémentaires : (1) la pixel-wise normalization dans le générateur pour éviter l’explosion des gradients ; (2) un bruit de type *minibatch standard deviation* injecté dans le discriminateur pour encourager la diversité des échantillons.

Résultats sur CelebA-HQ

L’expérimentation repose sur le dataset CelebA-HQ, une version haute définition (jusqu’à 1024×1024) du jeu de données de visages. Les images générées par ProGAN y atteignent un réalisme inédit pour l’époque, comme le montre la figure 6.8.



FIGURE 6.8 – Exemples de visages synthétiques générés par ProGAN sur le dataset CelebA-HQ, à une résolution de 1024×1024 pixels. Cette qualité visuelle est atteinte grâce à l’apprentissage progressif, qui permet au modèle de capturer d’abord la structure globale avant de raffiner les détails fins [10].

Limites et héritage

Malgré ses succès, ProGAN présente certaines limites : bien qu’il améliore nettement la stabilité, l’espace latent reste peu contrôlable, et les manipulations vectorielles y produisent des résultats parfois flous ou incohérents. De plus, la montée progressive reste rigide (résolutions fixées, schéma de transition imposé).

Cependant, l’impact de ProGAN est considérable. Il pose les fondations directes de StyleGAN, qui viendra enrichir la structure du générateur en apportant une séparation explicite du style à différents niveaux de granularité. ProGAN reste ainsi une étape majeure dans la montée en qualité des GANs.

6.2.3 StyleGAN et StyleGAN2 : séparation des styles et amélioration de la fidélité visuelle

À la suite de ProGAN, Tero Karras et al. [11] introduisent une refonte complète de l'architecture du générateur. Désormais centrée sur la notion de style. Le modèle StyleGAN repose sur l'idée d'un contrôle explicite du processus de génération à différentes échelles, inspiré des travaux en style transfer [36]. Cette nouvelle approche aboutit à des images plus réalistes, mais surtout à une meilleure maîtrise des facteurs de variation, comme la position, la coiffure ou les détails fins du visage.

Architecture : séparation explicite des styles

Contrairement aux GANs classiques où le vecteur latent z est directement injecté dans le réseau, StyleGAN introduit un réseau de mapping $f : \mathcal{Z} \rightarrow \mathcal{W}$ qui projette le bruit aléatoire dans un espace latent intermédiaire, plus apte à modéliser des facteurs de variation visuelle. À partir de ce vecteur $w \in \mathcal{W}$, le générateur applique, à chaque couche convulsive, une normalisation de type AdaIN, contrôlée par des styles appris.

Chaque couche est donc influencée par un style distinct, ce qui permet d'agir finement sur la génération à différentes résolutions. De plus, un bruit gaussien est injecté après chaque convolution, introduisant des variations stochastiques réalistes (taches de rousseur, brins de cheveux, etc.).

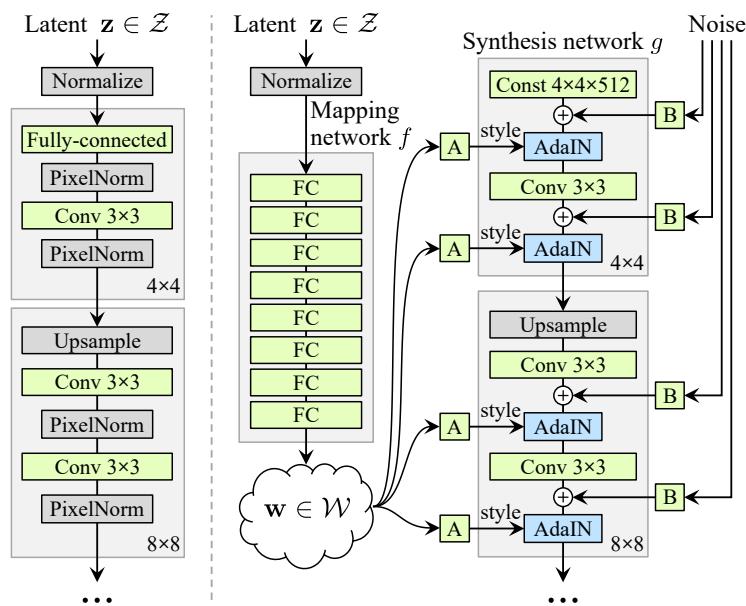


FIGURE 6.9 – Architecture du générateur de StyleGAN : un vecteur latent z est d'abord mappé vers un espace intermédiaire w , qui contrôle chaque couche via une normalisation adaptative. Un bruit stochastique est également injecté à chaque résolution pour enrichir les détails aléatoires [11].

Contrôle multi-échelle et mélange de styles

Cette séparation entre les couches permet d'exploiter un phénomène appelé *style mixing*, qui consiste à combiner les styles de deux vecteurs latents distincts. Par exemple, les couches responsables des basses fréquences (forme du visage, orientation) peuvent être influencées par un premier style, tandis que les couches plus fines (textures, cheveux, peau) proviennent d'un autre. Cela favorise la spécialisation des couches et améliore le désentrelacement des attributs.

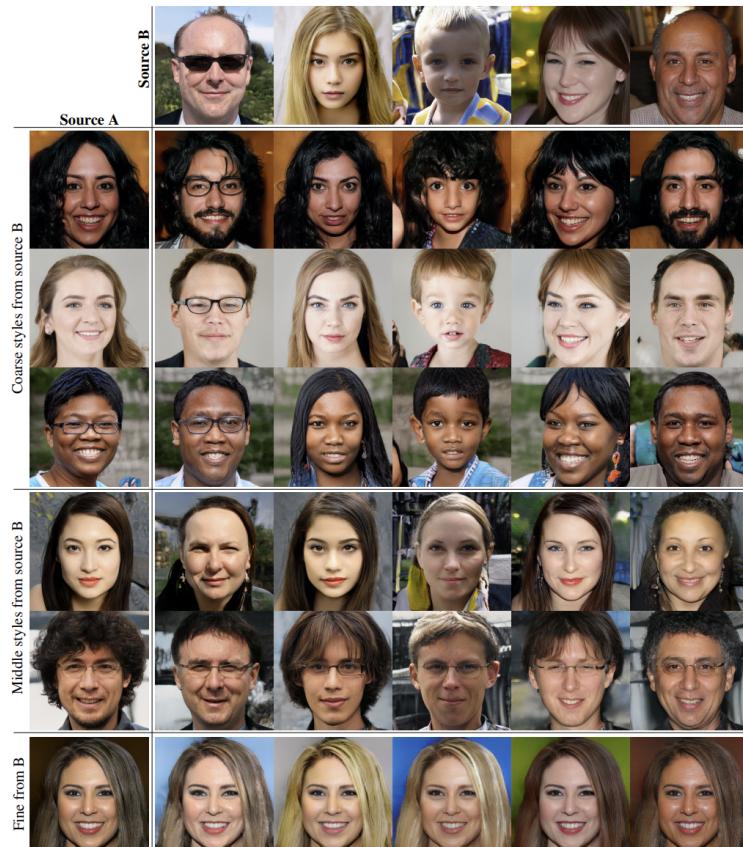


FIGURE 6.10 – Illustration du *style mixing* dans StyleGAN : chaque colonne résulte d'un mélange entre deux styles latents A et B à différentes résolutions (coarse, middle, fine). On observe que les couches profondes contrôlent la structure globale (pose, lunettes), tandis que les couches superficielles influencent les détails fins (peau, cheveux) [11].

Qualité visuelle et désentrelacement

StyleGAN permet d'atteindre des scores FID inégalés à l'époque, en particulier sur le nouveau dataset FFHQ, plus riche et varié que CelebA-HQ. Cette amélioration provient à la fois de la montée en résolution, rendue possible par la croissance progressive héritée de ProGAN, et de la réorganisation du pipeline génératif. En séparant les rôles du bruit aléatoire $z \in \mathcal{Z}$ et du style appris $w \in \mathcal{W}$, le générateur

devient plus sensible aux ajustements ciblés dans l'espace latent.

Karras et al. [11] introduisent deux métriques originales pour quantifier ce contrôle :

- la *perceptual path length* mesure la régularité des interpolations entre deux styles : plus cette longueur est faible, plus les interpolations sont continues et lisses au niveau perceptif,
- la *linear separability* quantifie la capacité à isoler des attributs visuels (ex. : sourire, sexe, lunettes) via des directions latentes linéaires dans \mathcal{W} .

Les résultats expérimentaux montrent que \mathcal{W} est significativement mieux désentrelacé que \mathcal{Z} , c'est-à-dire que chaque dimension tend à contrôler un facteur de variation indépendant. Cette propriété favorise les manipulations d'attributs sans interférence. Comme par exemple modifier la coupe de cheveux sans changer l'éclairage ou l'orientation du visage.

Cette nouvelle structuration latente confère à StyleGAN une expressivité bien supérieure à ses prédecesseurs, tant pour la génération que pour la réutilisation contrôlée dans des tâches de morphing ou d'édition faciale.

Améliorations apportées par StyleGAN2

Malgré ses réussites, StyleGAN présente certains artefacts structurels, comme des bulles ou motifs répétitifs qui apparaissent notamment sur les fonds ou dans les cheveux. Ces artefacts résultent en partie de la normalisation AdaIN, qui perturbe la distribution statistique des activations et contraint excessivement le signal.

Pour y remédier, Karras et al. [12] proposent StyleGAN2, une refonte subtile mais décisive de l'architecture. Ils suppriment AdaIN au profit d'une *weight demodulation*, une opération qui agit directement sur les poids des convolutions pour normaliser l'échelle des activations, sans modifier leurs statistiques. Ce changement permet d'éviter les corrélations artificielles entre canaux.



FIGURE 6.11 – Artefacts visibles dans StyleGAN, causés par la normalisation AdaIN. Ces distorsions, souvent en forme de bulles, apparaissent dès les premières couches de convolution (résolution 64×64) et sont présentes dans toutes les feature maps internes [12].

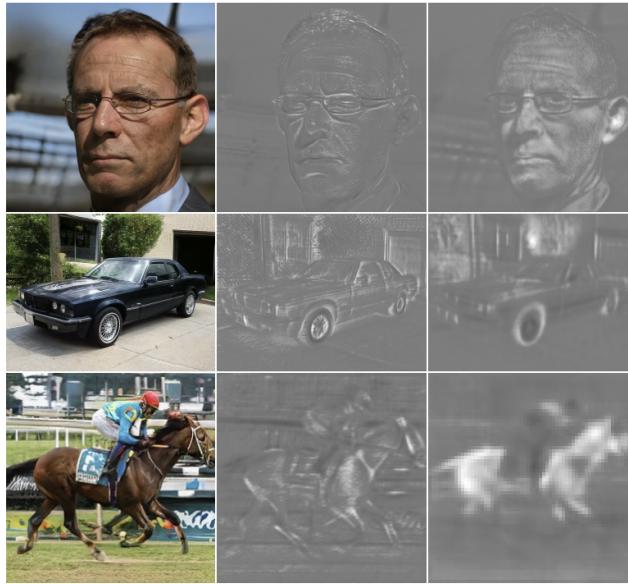


FIGURE 6.12 – Suppression des artefacts dans StyleGAN2 grâce à la technique de *weight demodulation*. Les activations deviennent plus homogènes et les images synthétisées présentent une qualité visuelle plus régulière [12].

StyleGAN2 introduit également :

- une meilleure initialisation des poids et une régularisation plus efficace des activations,
- une architecture de convolution redéfinie, plus stable à l'entraînement,
- une fusion plus cohérente entre bruit et signal latent, réduisant les artefacts contextuels.

Ces modifications permettent une nette amélioration des scores FID, en particulier à très haute résolution (1024×1024), et garantissent une synthèse d'image plus naturelle, avec moins d'effets indésirables et une meilleure généralisation sur des visages hors distribution.

Impact et postérité

Avec StyleGAN et StyleGAN2, Karras et al. [11] [12] franchissent un seuil critique dans la génération réaliste de visages humains. Ces architectures deviennent la base de nombreuses variantes ultérieures (StyleGAN3, InterFaceGAN, StyleCLIP, etc.) et constituent désormais un standard dans la synthèse d'images haute qualité. La distinction entre attributs stochastiques et contrôlables, l'émergence d'un espace latent manipulable, et l'introduction de styles injectables couche par couche définissent un tournant majeur dans la conception des générateurs GAN.

6.3 Capacités de génération et détournements frauduleux

L'évolution des GAN ne se limite pas à l'amélioration visuelle des images générées. Ces modèles possèdent aujourd'hui des capacités expressives étendues qui les rendent à la fois plus impressionnantes et plus redoutables lorsqu'ils sont utilisés à des fins malveillantes. Leur flexibilité permet non seulement de synthétiser des images, mais aussi de manipuler des identités, de modifier des vidéos ou de falsifier des documents avec un réalisme troublant.

L'une des propriétés fondamentales exploitées dans les GANs est la continuité de l'espace latent. Deux vecteurs latents z_1 et z_2 peuvent être interpolés linéairement, produisant une séquence fluide d'images intermédiaires. Ce comportement, signe d'une organisation sémantique cohérente dans l'espace latent, est parfois utilisé pour manipuler progressivement des attributs visuels d'un visage de manière contrôlée. Les GANs conditionnels permettent de générer des contenus guidés par des attributs explicites : âge, sexe, expression faciale ou même description textuelle. Cette capacité ouvre la voie à des manipulations ciblées, comme la génération de visages réalistes répondant à des critères précis, parfois observée dans des contextes de désinformation ou d'usurpation. En contrepartie, le conditionnement imparfait peut parfois laisser des artefacts exploitables pour la détection.

Les architectures modernes permettent aussi la génération de séquences vidéo, par animation faciale ou synthèse d'expressions. En partant d'une seule image fixe, certains GANs sont capables de produire des mouvements de bouche ou des expressions crédibles, créant l'illusion que la personne parle réellement. Ces techniques, souvent désignées sous le terme de deepfake, soulèvent des préoccupations quant à la véracité des contenus audiovisuels. Notamment lorsqu'elles sont utilisées pour manipuler des discours politiques ou diffuser des propos mensongers.

Ces avancées technologiques ont mené à plusieurs détournements concrets. Des visages artificiels sont utilisés pour créer de faux profils crédibles en ligne, parfois dans le but d'arnaquer, d'espionner ou de propager de fausses informations. Des photos d'identité générées par GAN peuvent servir à contourner des systèmes de vérification biométrique. Plus largement, des plateformes clandestines proposent à la vente des modèles pré-entraînés ou des services de génération sur commande, contribuant à l'essor d'un marché noir de la falsification numérique.

Face à ces usages malveillants, il devient essentiel de développer des méthodes robustes et automatisées de détection. L'étude des propriétés propres aux images générées par des GANs (interpolations, conditionnement, artefacts subtils), constituent autant d'indices exploitables pour la détection.

Chapitre 7

Méthodes de détection des images générées par GAN

7.1 Méthodes basées sur des descripteurs et artefacts visuels

7.1.1 Analyse de matrices de co-occurrence

L’approche proposée par Nataraj et al. [13] consiste à exploiter les matrice de co-occurrence extraites directement dans le domaine pixel, sans recours à des résidus d’image ni à des filtres préalables, ce qui constitue une différence fondamentale avec les techniques de stéganalyse classiques [37, 38]. Ces matrices sont calculées séparément sur chacun des trois canaux (R , G , B), formant un tenseur d’entrée de taille $3 \times 256 \times 256$ qui est ensuite injecté dans un réseau convolutif profond.

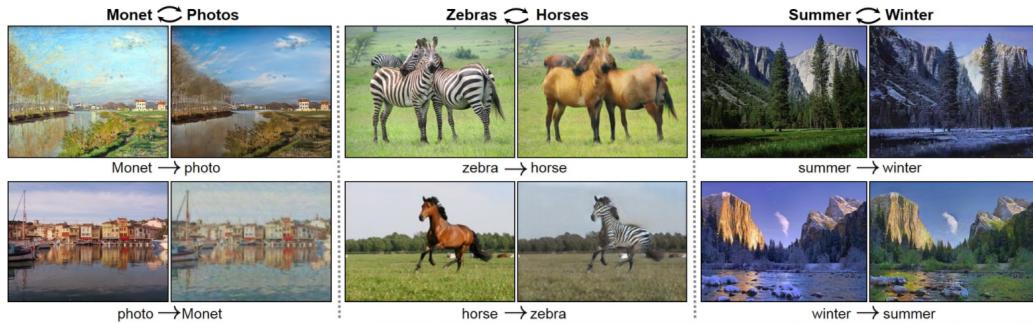


FIGURE 7.1 – Exemples d’images générées par CycleGAN et StarGAN, extraits de [13].

Principe des matrices de co-occurrence

Dans le contexte du traitement d’images, une matrice de co-occurrence $P(i, j)$ mesure la fréquence conjointe d’occurrence de deux niveaux de gris i et j dans une position relative donnée (par exemple, entre un pixel et son voisin horizontal direct). Mathématiquement, on la définit par :

$$P(i, j) = \sum_{x,y} \delta(I(x, y) = i \wedge I(x + \Delta x, y + \Delta y) = j) \quad (7.1)$$

où $I(x, y)$ est l’intensité du pixel à la position (x, y) , $(\Delta x, \Delta y)$ représente le décalage directionnel (typiquement $(1, 0)$ pour le voisin de droite), et δ est la fonction indicatrice. La normalisation de cette matrice donne une distribution empirique des co-occurrences.

Les images naturelles présentent des régularités locales, corrélations entre pixels voisins, qui sont perturbées par les processus de génération des GAN. L’analyse directe des co-occurrences sur les canaux (R, G, B) permet donc de capturer ces anomalies sans appliquer de filtrage ou de calcul de résidu, ce qui simplifie le processus de détection tout en exploitant la richesse des statistiques locales.

Architecture de détection proposée

Plutôt que d’extraire des caractéristiques manuelles sur les co-occurrences, Nataraj et al. font passer directement les matrices dans un réseau convolutif. Chaque image est convertie en un tenseur $T \in \mathbb{R}^{3 \times 256 \times 256}$, formé par les co-occurrences calculées indépendamment sur chaque canal (R, G, B).

L’architecture utilisée est la suivante :

- **Étape 1** : convolution 3×3 avec 32 filtres + ReLU
- **Étape 2** : convolution 5×5 avec 32 filtres + max pooling

- **Étape 3** : convolution 3×3 avec 64 filtres + ReLU
- **Étape 4** : convolution 5×5 avec 64 filtres + max pooling
- **Étape 5** : convolution 3×3 avec 128 filtres + ReLU
- **Étape 6** : convolution 5×5 avec 128 filtres + max pooling
- **Dense** : deux couches entièrement connectées de 256 neurones
- **Sortie** : activation sigmoïde pour la classification binaire

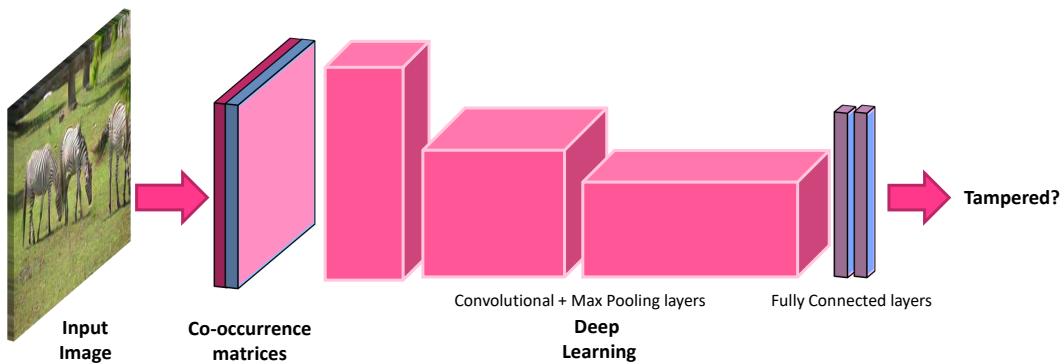


FIGURE 7.2 – Schéma de l'architecture complète proposée par Nataraj et al., combinant matrices de co-occurrence RGB et CNN [13].

Évaluation et robustesse

L'approche a été évaluée sur deux jeux de données synthétiques :

- **CycleGAN** [39] : traductions d'images (été→hiver, cheval→zèbre, etc.), 36 302 images.
- **StarGAN** [40] : visages de célébrités modifiés (attributs comme âge, sexe, expression), 19 990 images.

Les résultats obtenus indiquent une précision de 99.71% sur CycleGAN et 99.37% sur StarGAN. En croisant les jeux (entraînement sur un, test sur l'autre), les performances restent élevées, confirmant la généralisation du modèle.

TABLE 7.1 – Expérience de généralisation croisée entre jeux de données

Entraînement	Test	Précision
CycleGAN	StarGAN	99.49%
StarGAN	CycleGAN	93.42%

Une étude de robustesse face à la compression JPEG révèle que l'algorithme est sensible à ce type de déformation lorsqu'il n'est pas entraîné dessus, mais qu'il

conserve une bonne performance si les données compressées sont incluses dès l'apprentissage.

TABLE 7.2 – Effet de la compression JPEG sur la performance de détection

Qualité JPEG	Train : original	Train : compressé
QF = 95	74.50%	93.78%
QF = 85	69.46%	91.61%
QF = 75	64.46%	87.31%

TABLE 7.3 – Comparaison des précisions de classification sur CycleGAN par catégorie

Méthode	Steganalysis	Cozzolino2017	XceptionNet	Proposée
Précision moyenne (%)	94.40	95.07	94.49	97.84

En comparaison avec des méthodes de pointe comme XceptionNet, les modèles à base de résidus [41], ou encore des méthodes issues de la stéganalyse, la méthode proposée obtient la meilleure précision moyenne. Cette méthode illustre comment un simple indice statistique, la co-occurrence pixel à pixel, peut devenir un puissant outil de détection lorsqu'il est intégré à une architecture d'apprentissage profond, en exploitant les traces subtiles laissées par les processus de génération artificielle. Elle présente un bon compromis entre simplicité de traitement (aucune opération de filtrage ou de résidu n'est nécessaire) et efficacité de classification. Toutefois, sa sensibilité aux déformations comme la compression JPEG ou la résolution variable des images pourrait constituer une limite dans des scénarios plus réalistes ou adversariaux. Il serait pertinent d'envisager des méthodes complémentaires s'appuyant sur la modélisation ou la simulation explicite d'artefacts, comme nous le verrons dans la section 7.1.3.

7.1.2 Analyse colorimétrique

Dans leur étude, McCloskey et Albright [14] proposent une approche originale pour détecter les images produites par des GANs en s'appuyant non pas sur les artefacts visuels apparents, mais sur une analyse approfondie du traitement de la couleur effectué dans l'architecture même du générateur. Ils identifient deux indices principaux distinctifs par rapport aux caméras physiques : le mécanisme de formation des canaux RGB et la distribution des pixels saturés, influencée par les normalisations internes au réseau.

Formation des couleurs : une analogie défaillante

Le dernier bloc du générateur GAN prend une représentation en profondeur, composée de K couches de feature maps ($K > 3$), et la projette sur trois canaux couleur RGB à l'aide d'une convolution 1×1 sur les canaux, comme illustré dans la Figure 7.3. Ce mécanisme est analogue, dans sa structure, à la réponse spectrale d'un capteur photo à travers un filtre de Bayer. Toutefois, les poids appris dans un GAN ne sont soumis à aucune contrainte physique : ils peuvent se chevaucher fortement, être négatifs, ou corrélés, contrairement aux filtres optiques réels.

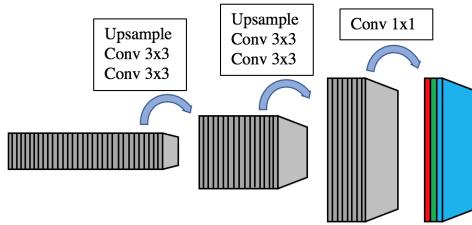


FIGURE 7.3 – Architecture simplifiée du générateur (tirée de [14]). La conversion des couches profondes vers la sortie RGB est réalisée par convolution 1×1 sur les canaux.

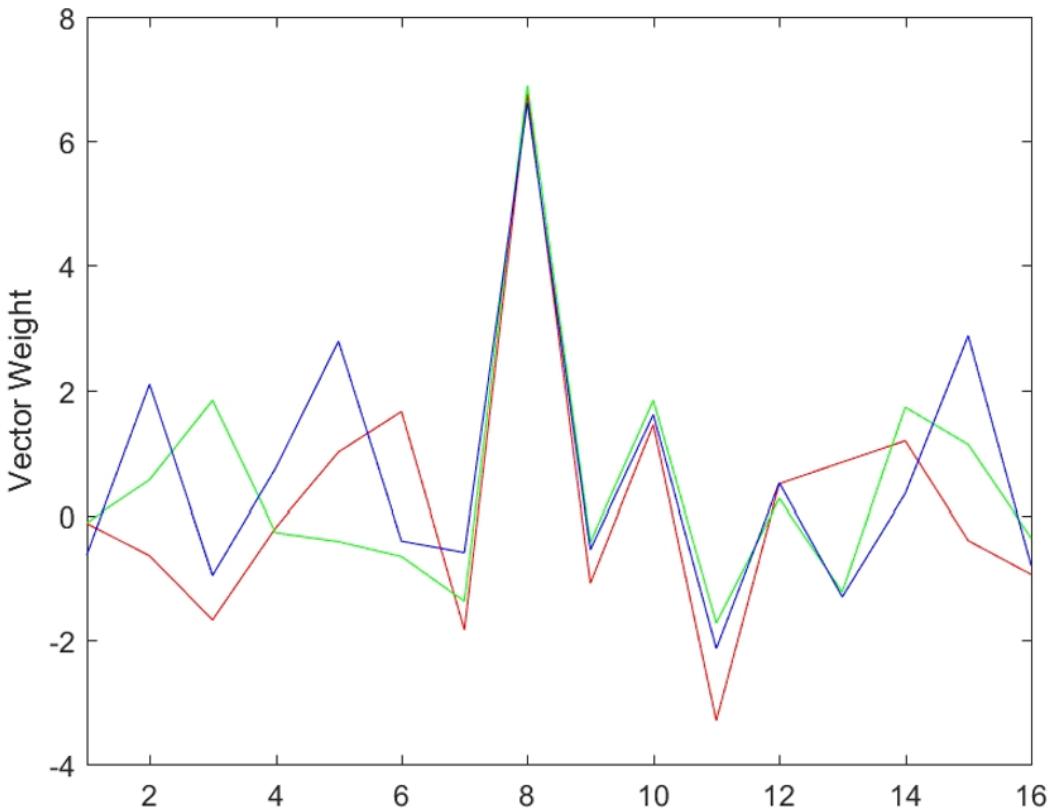


FIGURE 7.4 – Poids de projection utilisés par le générateur GAN pour produire les canaux R, G, B à partir de 16 couches internes. La forte redondance et corrélation entre les trois courbes contraste avec les profils spectraux d'un capteur réel.

Les auteurs montrent que les poids appris par le réseau (Figure 7.4) présentent une forte similarité entre canaux, ce qui induit une corrélation inhabituelle entre les composantes chromatiques. Pour l'évaluer, ils projettent les images dans l'espace de chromaticité rg, où :

$$r = \frac{R}{R + G + B}, \quad g = \frac{G}{R + G + B} \quad (7.2)$$

L'idée est que les images GAN auront une densité plus concentrée dans l'espace rg, traduisant un spectre chromatique plus uniforme, dû à la projection redondante.

Fréquence des pixels saturés : effet de la normalisation

La seconde caractéristique révélatrice d'une image GAN est liée à la rareté des pixels saturés. Contrairement aux images naturelles, qui comportent fréquemment des zones sur-exposées (ou sous-exposées), les générateurs GAN appliquent des normalisations internes pour stabiliser l'apprentissage, qui contraignent la dynamique des sorties.

Dans l'implémentation de [10], la normalisation est effectuée pixel par pixel après chaque couche convolutionnelle, selon :

$$b_{j,x,y} = \frac{a_{j,x,y}}{\sqrt{\frac{1}{N} \sum_{c=0}^{N-1} (a_{c,x,y})^2 + \epsilon}} \quad (7.3)$$

où N est le nombre de feature maps, et a la valeur brute avant normalisation. Cette opération homogénéise les valeurs dans une plage régulière, empêchant les saturations classiques observées dans les images HDR capturées par caméra.

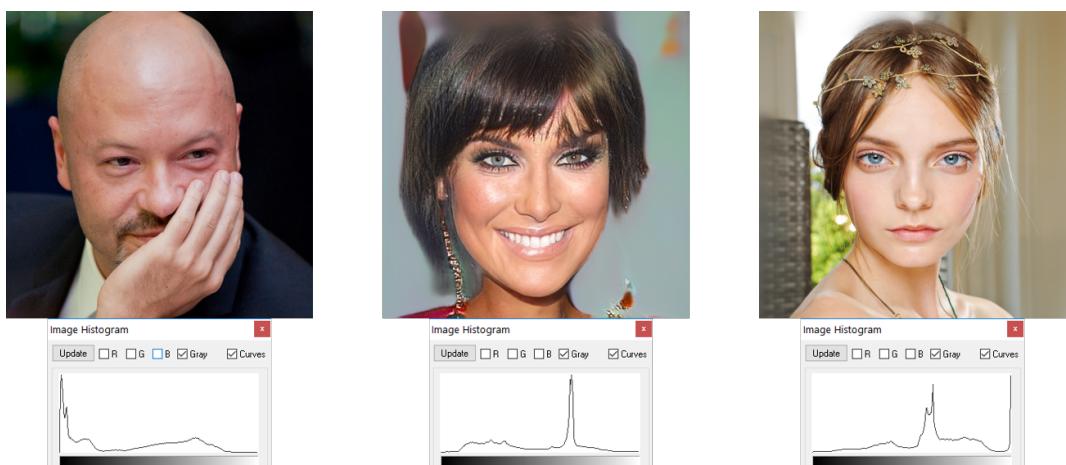


FIGURE 7.5 – Comparaison visuelle et histogrammes : les images GAN (centre) tendent à éviter la saturation, contrairement aux images réelles qui présentent des plages surexposées (droite) ou sous-exposées (gauche) [14].

Deux détecteurs complémentaires

Les auteurs proposent deux méthodes distinctes pour tirer parti de ces observations :

- Chromaticité rg : une représentation des images sous forme d'histogrammes bivariés (r, g), classifiés par un modèle pré-entraîné de type INH-VGG (adapté de [42]).
- Saturation : un vecteur de caractéristiques basé sur la fréquence des pixels proches de la saturation (intensité $\geq 240, 245, 250, 255$) ou de l'obscurcissement (≤ 15), classifié via un SVM linéaire.

Ces deux détecteurs sont évalués séparément sur deux sous-ensembles du Media Forensics Challenge (MFC18) :

- GAN Crop : patchs purement générés par GAN.
- GAN Full : visages GAN insérés dans une scène réelle.

Résultats expérimentaux

Le détecteur basé sur la saturation obtient de bons résultats, notamment sur les images entièrement synthétiques ($AUC = 0.70$). Il reste modérément efficace sur les images partiellement manipulées ($AUC = 0.61$), bien que la zone GAN soit plus petite.

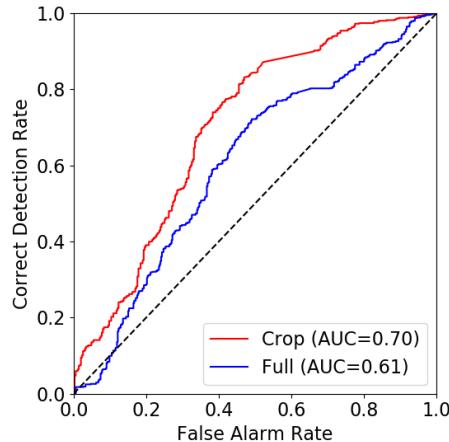


FIGURE 7.6 – Courbes ROC pour le détecteur basé sur la saturation : bons résultats sur GAN Crop, performance réduite sur GAN Full.

En revanche, le détecteur basé sur la chromaticité rg s'avère peu discriminant (AUC proches de 0.55), ce que les auteurs attribuent à un entraînement insuffisant du modèle sur les images GAN, et à la présence de flous ou retouches dans certaines images réelles.

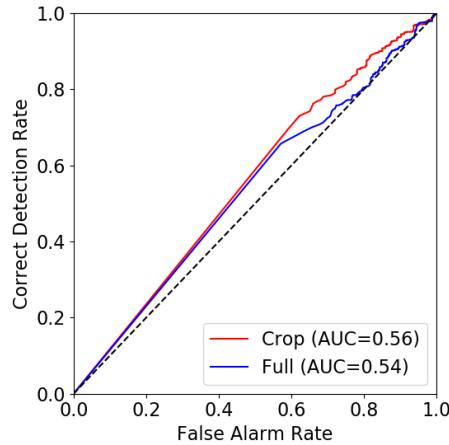


FIGURE 7.7 – Courbes ROC du détecteur de chromaticité : faible pouvoir discriminant.

Discussion

Cette approche se distingue par son positionnement hybride : elle ne s'appuie pas sur l'apprentissage bout-à-bout, mais sur des indices structurels propres au fonctionnement des GANs. Le fait d'identifier que la normalisation interne et le mécanisme de synthèse couleur altèrent statistiquement les images produites constitue une avancée conceptuelle importante.

La méthode par saturation présente une efficacité significative avec une complexité faible, et ne dépend pas de l'architecture GAN. Elle pourrait toutefois être contournée par un post-traitement intelligent réintroduisant des saturations artificielles. En revanche, l'analyse colorimétrique nécessite des modèles plus robustes pour atteindre un niveau exploitable.

7.1.3 Simulation d'artefacts

Artéfacts typiques et motivation de la simulation

L'un des traits caractéristiques des images produites par les GAN est la présence d'artefacts visuels subtils, souvent perceptibles par des motifs répétitifs, des discontinuités de textures ou encore des incohérences structurelles. Ces artefacts ne sont pas injectés de manière explicite par le générateur mais émergent du processus d'entraînement. En particulier dans les zones à fort contraste, aux jonctions de régions sémantiques ou dans des détails fins (les pupilles, les dents, ou les cheveux).

Plutôt que de chercher à détecter passivement ces anomalies comme des défauts visuels a posteriori, Zhang et al. [15] proposent une approche originale consistant à simuler ces artefacts pour en faire un signal supervisé d'apprentissage. L'intuition

est la suivante : si l'on parvient à modéliser les défauts caractéristiques des images GAN, on peut ensuite entraîner un détecteur explicite capable d'apprendre à les reconnaître, même lorsqu'ils sont subtils ou partiellement atténus.

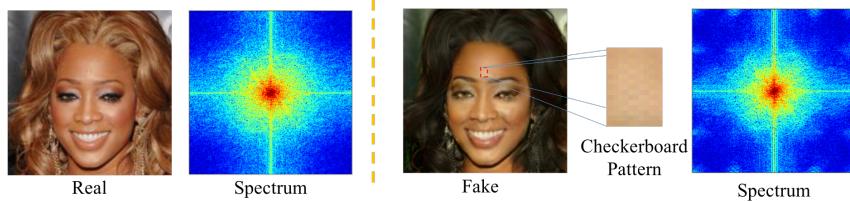


FIGURE 7.8 – Exemple d'artefact typique introduit par les GAN : motif périodique de type « damier », causé par les opérations de déconvolution [15].

Analyse fréquentielle des artefacts GAN

Au-delà des défauts visibles localement (flous, textures incohérentes), Zhang et al. [15] mettent en évidence une signature plus globale, observable dans le domaine fréquentiel. En particulier, les images synthétiques présentent souvent des régularités spatiales anormales, sous la forme de motifs périodiques ou de pics localisés dans certaines bandes de fréquences.

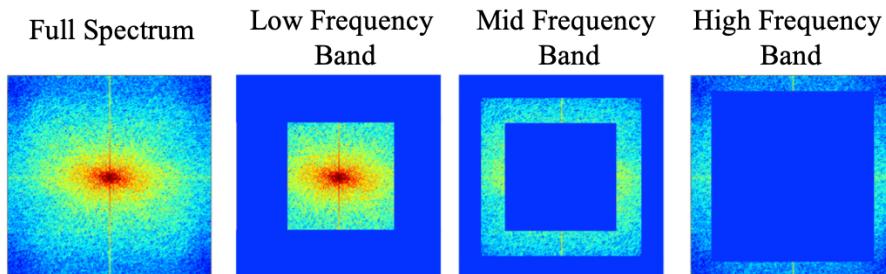


FIGURE 7.9 – Spectre de fréquence moyen d'un grand ensemble d'images GAN. On observe des pics anormaux dans certaines bandes spécifiques, absents dans les images naturelles [15].

Ces irrégularités sont typiquement causées par les opérations de déconvolution utilisées pour l'upsampling dans de nombreux générateurs GAN. Lorsqu'elles sont mal paramétrées, ces convolutions transposées génèrent des artefacts périodiques, visibles aussi bien à l'oeil nu qu'en transformée de Fourier.

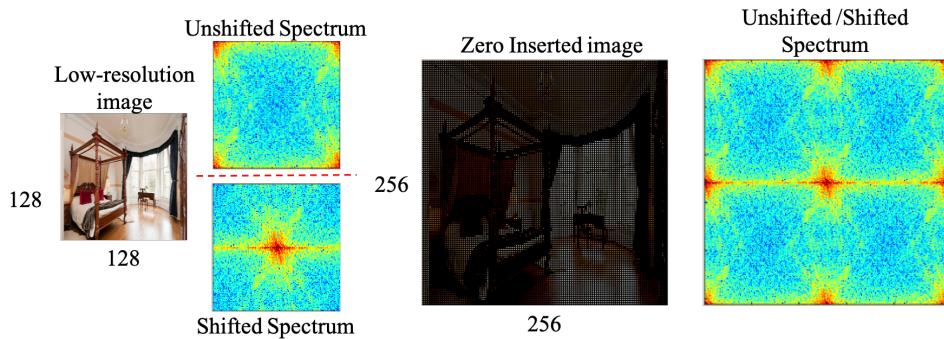


FIGURE 7.10 – Comparaison entre le spectre fréquentiel d'une image GAN (gauche) et d'une image réelle (droite). Les images GAN présentent des structures périodiques fortement localisées, absentes dans les images naturelles [15].

Cette analyse spectrale vient justifier la conception du simulateur, qui vise notamment à reproduire ce type de signature fréquencelle de manière contrôlée dans les images d'entraînement.

Réseau de simulation d'artefacts GAN

Cette approche repose sur l'architecture appelée AutoGAN. Il s'agit d'un réseau convolutionnel léger conçu pour appliquer artificiellement des distorsions similaires à celles que l'on observe dans des images synthétiques, tout en préservant les grandes structures sémantiques de l'image.

Le simulateur prend une image réelle I en entrée, et génère une image $\tilde{I} = \mathcal{A}(I)$ contenant des artefacts réalistes. Pour forcer la similitude avec les vraies images GAN, les auteurs introduisent une perte adversarielle où le simulateur est opposé à un discriminateur D , chargé de distinguer les images simulées des images réelles. Cela pousse le simulateur à générer des artefacts réalistes tout en conservant la structure globale de l'image.

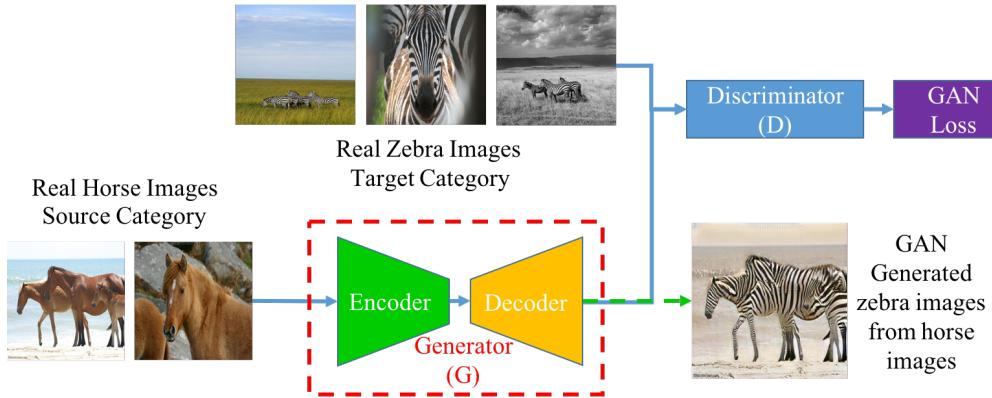


FIGURE 7.11 – Pipeline complet de la méthode : le simulateur applique des artefacts artificiels, et le détecteur est entraîné à distinguer images réelles, images simulées, et images GAN [15].

La fonction de perte combinée est la suivante :

$$L = \sum_{i=1}^n \log(D(I_i)) + \log(1 - D(G(I_i))) + \lambda \|I_i - G(I_i)\|_1$$

où D est le discriminateur, G le générateur (simulateur), et λ un coefficient de pondération. Cette formulation encourage la génération d'artefacts subtils tout en maintenant une forte ressemblance visuelle avec les images d'origine.

Apprentissage et détection supervisée

Une fois le simulateur entraîné, il permet de générer un grand nombre d'images synthétiques contaminées artificiellement, sans recourir à des jeux de données GAN coûteux à générer. Ces images servent ensuite à entraîner un réseau convolutif résiduel (ResNet-50), qui apprend à distinguer les images réelles des images modifiées (qu'elles proviennent du simulateur ou d'un GAN véritable).

Ce détecteur est donc entraîné sur :

- des images réelles (issues d'ImageNet),
- des images modifiées par le simulateur AutoGAN,
- et des images GAN (principalement issues de CycleGAN ou StarGAN).

L'avantage de ce schéma est double : il permet un enrichissement des données d'entraînement, et offre une meilleure robustesse aux variantes de GAN jamais vues durant l'apprentissage.

Résultats expérimentaux et robustesse

L'évaluation est menée sur plusieurs jeux de données synthétiques, en testant la capacité du modèle à généraliser à des architectures GAN non vues à l'entraînement. Par exemple, un entraînement uniquement basé sur des images simulées à partir de CycleGAN (ou d'images naturelles via AutoGAN), permet d'atteindre une précision supérieure à 94 % sur StarGAN, une architecture pourtant non vue pendant l'entraînement.

Les expériences présentées montrent que l'utilisation du simulateur AutoGAN combinée à une représentation fréquentielle des images améliore significativement la capacité du détecteur à généraliser. En particulier, les classificateurs entraînés avec AutoGAN et testés sur des images issues de CycleGAN ou StarGAN atteignent respectivement des précisions allant jusqu'à 94,0 % et 98,7 %. En comparaison, les méthodes classiques fondées sur les pixels peinent à dépasser les 65–70 % de précision lorsque le domaine source diffère du domaine cible. Ces résultats valident l'hypothèse selon laquelle les artefacts d'up-sampling, partagés entre différentes familles de GAN, peuvent être capturés en amont via une simulation fidèle dans le domaine fréquentiel.

En plus d'une meilleure précision, l'approche démontre une capacité à localiser spatialement les artefacts dans les images GAN, notamment en activant fortement certaines zones suspectes (yeux, bouches, textures de fond).

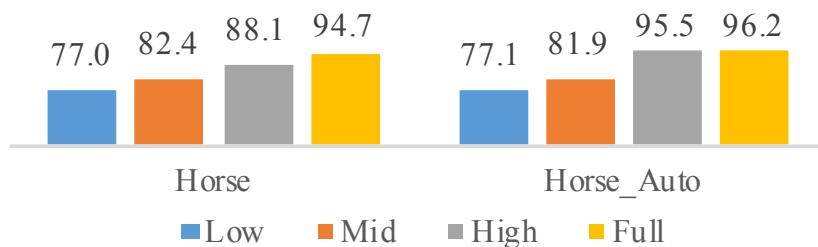


FIGURE 7.12 – Analyse fréquentielle des artefacts simulés : les distorsions introduites reproduisent des schémas proches de ceux observés dans les GAN réels [15].

Les histogrammes comparent l'énergie spectrale (après décomposition en bande de fréquences spatiales) entre une image GAN (Horse) et une image réelle altérée par le simulateur (Horse_Auto). La similarité des distributions valide la capacité du simulateur à reproduire les signatures fréquentielles typiques des artefacts GAN.

Cette approche propose une innovation méthodologique forte en transposant le paradigme de la data augmentation à la détection de GAN, via la simulation contrôlée d'artefacts caractéristiques. Cela permet non seulement de renforcer la généralisation du détecteur, mais aussi de gagner en explicabilité, en liant les décisions à des indices visuels précis.

7.2 Méthodes fondées sur l'apprentissage automatique

7.2.1 Détection par CNN et SVM

Du descripteur SRM au CNN constraint

Certaines des premières méthodes efficaces de détection d'images manipulées reposent sur l'utilisation de descripteurs fondés sur des résidus de bruit haute-fréquence inspirés des modèles de stéganalyse. Le Spatial Rich Model (SRM) et le Subtractive Pixel Adjacency Matrix (SPAM) en sont deux exemples notables, utilisant des opérations de filtrage, quantification scalaire, co-occurrences et histogrammes pour capturer les microstructures statistiques altérées lors de manipulations [16].

Cozzolino et al. proposent une contribution originale en montrant que ces descripteurs peuvent être reformulés comme un réseau convolutif constraint : chaque étape (filtrage, quantification, histogramme), à base de calculs locaux, peut être implémentée comme une couche convolutionnelle à poids figés. Le passage du modèle SRM à un réseau CNN complet est détaillé dans plusieurs schémas, dont la Figure 7.13 qui illustre l'intégration finale des couches dans une architecture de classification supervisée.

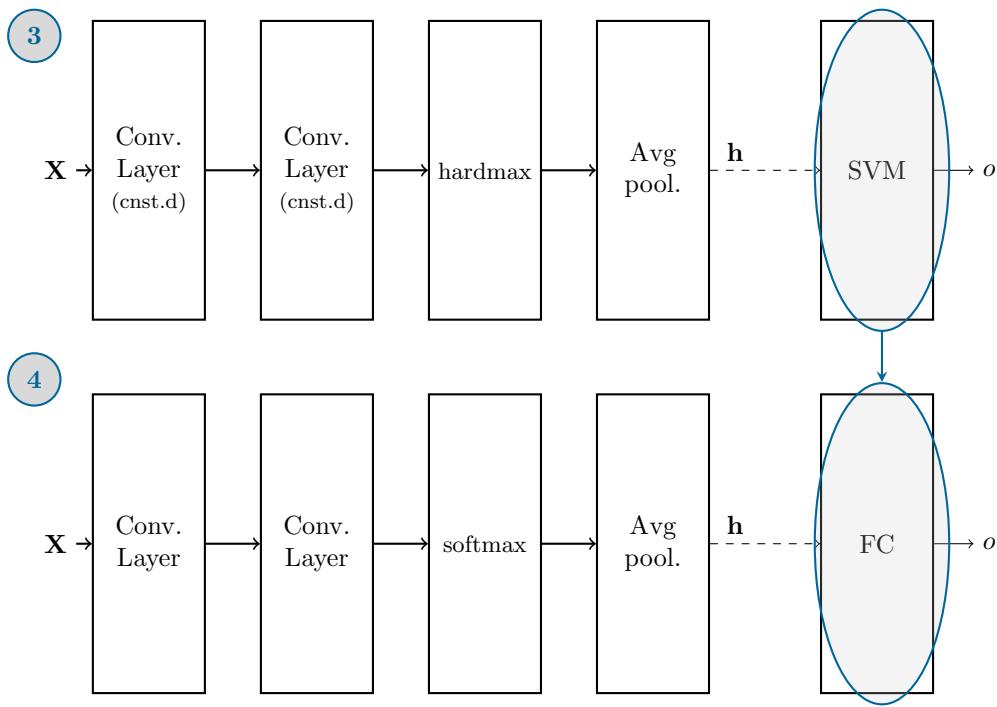


FIGURE 7.13 – Conversion du pipeline SRM vers un CNN contraint avec couches convolutionnelles et moyenne spatiale [16].

Une fois cette "équivalence" mise en évidence, les auteurs suppriment progressivement les contraintes : les poids du filtrage sont initialisés selon le SRM, mais ensuite ajustés par un entraînement supervisé sur un ensemble d'images altérées. Le passage du hardmax au softmax permet le calcul de gradients pour l'apprentissage. L'ensemble est alors optimisé conjointement via descente de gradient stochastique.

Comparaison aux approches SVM classiques

Cette réinterprétation des descripteurs comme CNN permet d'aller au-delà du couple SRM + SVM traditionnel, souvent utilisé pour classifier les histogrammes issus des cooccurrences. En effet, une fois reformulé comme un réseau de neurones convolutionnel contraint, le modèle peut être entraîné de bout en bout sur un jeu d'images manipulées, ce qui permet une optimisation conjointe de l'extraction de caractéristiques et de la classification.

Les résultats expérimentaux présentés par Cozzolino et al. mettent en évidence que le CNN ainsi appris surpassé les performances du SVM, notamment lorsque la taille du jeu d'entraînement est limitée. L'approche CNN bénéficie en effet d'une initialisation fondée sur les filtres SRM, puis d'un affinement supervisé via rétropropagation. Ce gain de performance est particulièrement visible pour des manipulations subtiles comme les floutages faibles ou les légers redimensionnements.

La Figure 7.14 présente deux exemples issus des expérimentations : en haut, un cas de splicing flouté ; en bas, un exemple de copy-move avec redimensionnement. Dans les deux cas, la comparaison entre la carte de chaleur produite par SRM+SVM et celle issue du CNN met en évidence une localisation plus précise, moins bruitée, et mieux centrée sur les zones effectivement modifiées, confirmant l'intérêt de l'approche par CNN dans ce contexte.

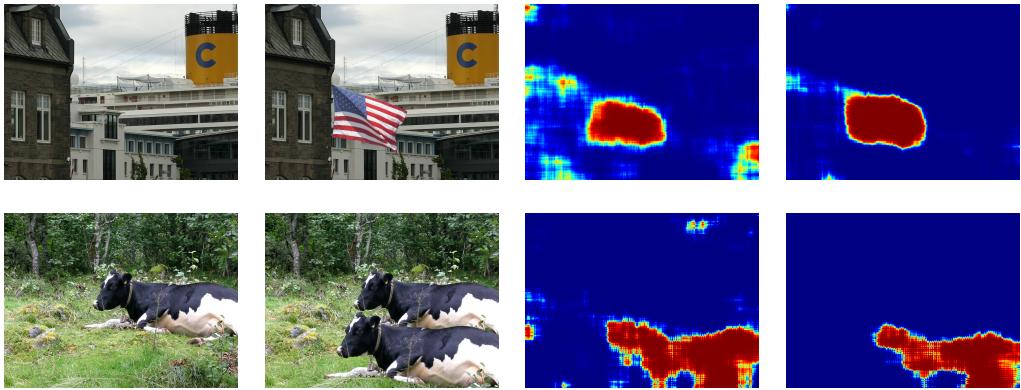


FIGURE 7.14 – Comparaison de la localisation des zones falsifiées pour deux types de manipulations : splicing flouté (en haut) et copy-move redimensionné (en bas). De gauche à droite : image originale, image falsifiée, carte de chaleur obtenue avec SRM+SVM, carte obtenue avec le CNN proposé [16].

Empreintes résiduelles et identification de source

Dans une démarche connexe, Marra et al. proposent d'utiliser des CNN pour exploiter les "empreintes artificielles" laissées par les GAN dans leurs images générées. Le principe consiste à extraire des résidus de bruit $R_i = X_i - f(X_i)$, puis à estimer une moyenne \hat{F} , interprétée comme une empreinte stable du GAN utilisé. La Figure 7.15 illustre l'effet de l'accumulation sur cette estimation.

Les auteurs montrent que les images issues d'un même GAN corrèlent fortement avec leur propre empreinte, et non avec celles d'autres GAN, permettant une attribution fiable. Cela ouvre la voie à des méthodes supervisées ou semi-supervisées basées sur la reconnaissance d'empreintes GAN en entrée d'un CNN ou en post-traitement des résidus.

Discussion

La portée de ces approches est double : elles montrent que même des CNN simples initialisés par des descripteurs statistiques classiques, peuvent être très efficaces. Et elles mettent en évidence le potentiel des traces laissées par les GAN eux-mêmes, exploitables comme signature pour la détection. Dans les deux cas, la combinaison

entre des connaissances préalables (SRM, résidus) et l'apprentissage profond permet des gains notables de performance tout en conservant une explicabilité partielle du modèle.

7.2.2 Méthodes par empreinte GAN (fingerprints)

Nous avons déjà mentionné brièvement, dans la sous-section précédente, le travail de Marra et al. [17] qui repose sur l'hypothèse selon laquelle chaque GAN laisse une trace spécifique, une « empreinte numérique », dans les images qu'il génère. Nous détaillons ici cette approche fondée sur l'analyse de résidus et l'attribution de source.

L'étude commence par l'observation que, de la même manière que les capteurs d'appareils photo laissent une empreinte (le PRNU), les GAN introduisent également des patterns déterministes. En extrayant un grand nombre de résidus R_i à partir des images synthétiques, puis en les moyennant, les auteurs mettent en évidence une structure stable propre à chaque modèle génératif. La Figure 7.15 illustre l'effet du nombre de résidus N sur la stabilité de l'empreinte estimée.

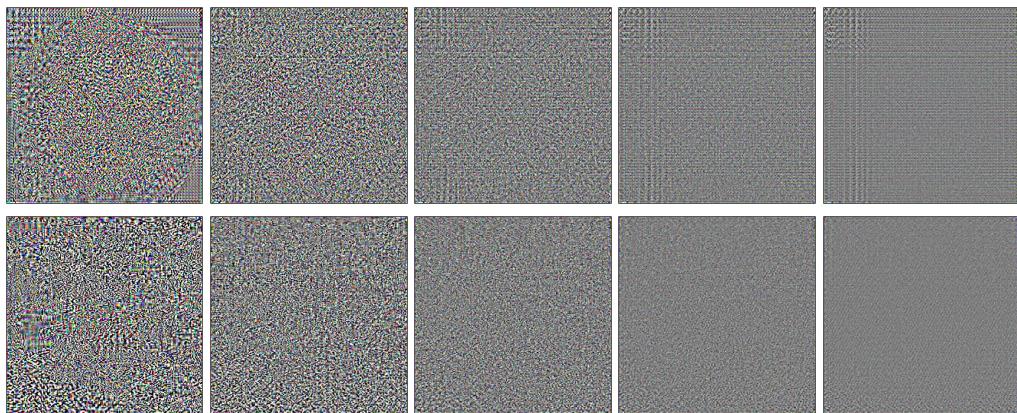


FIGURE 7.15 – Empreintes estimées avec $N = 2, 8, 32, 128, 512$ résidus pour CycleGAN (haut) et ProGAN (bas) [17].

Lorsque N augmente, le bruit aléatoire s'annule progressivement et un motif quasi-périodique récurrent apparaît. L'énergie de l'empreinte estimée suit une loi décroissante de la forme :

$$\hat{E}(N) = E_\infty + E_0 \cdot 2^{-N}$$

confirmant le caractère déterministe du motif extrait. L'étude de l'autocorrélation, illustrée en Figure 7.16, révèle une organisation interne non aléatoire.

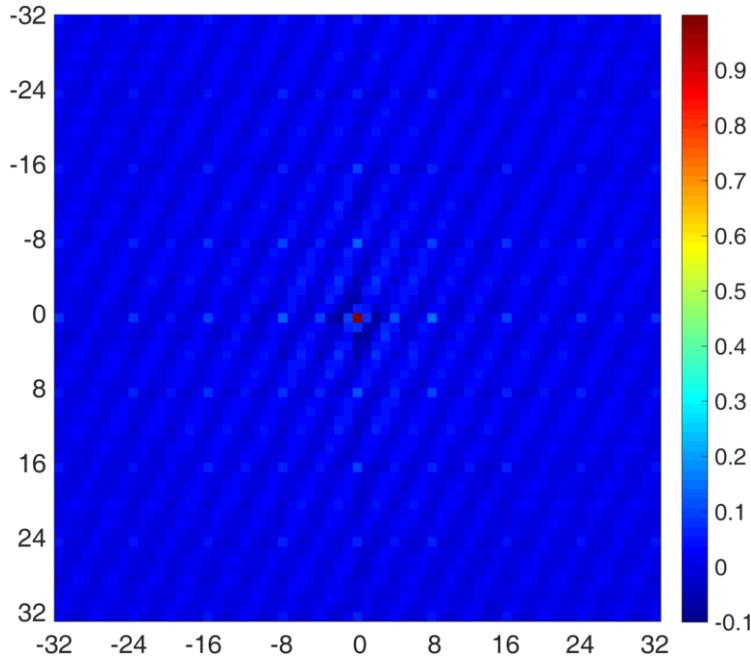


FIGURE 7.16 – Autocorrélation de l'empreinte estimée pour ProGAN ($N = 512$) [17].

Marra et al. poursuivent avec une expérience d'attribution de source. En comparant les corrélations des résidus d'une image avec différentes empreintes, ils démontrent que les images générées par un même GAN présentent une forte corrélation avec leur propre empreinte, et non avec celles des autres. La Figure 7.17 synthétise cette observation, illustrant des distributions bien séparées entre corrélation intra- et inter-GAN.

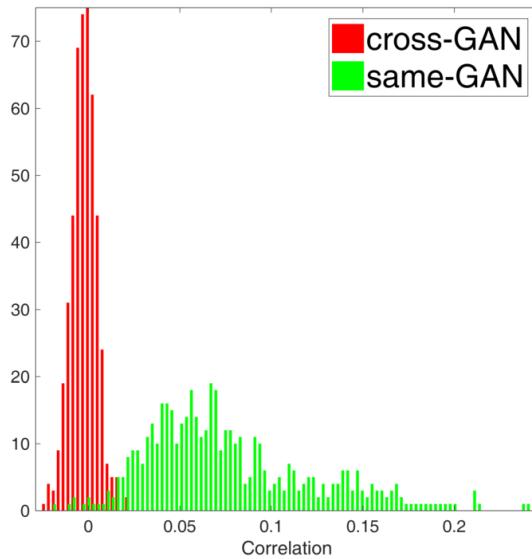


FIGURE 7.17 – Distribution des corrélations entre résidus d'image et empreintes GAN : forte distinction entre empreintes croisées et correspondantes [17].

Les auteurs étendent ensuite leur analyse à 20 modèles différents (CycleGAN, ProGAN, StarGAN) et deux appareils photo (Nikon-D90 et D7000). La matrice de corrélation moyenne présentée en Figure 7.18 révèle une forte diagonale dominante, signe que chaque empreinte est spécifique à son générateur.

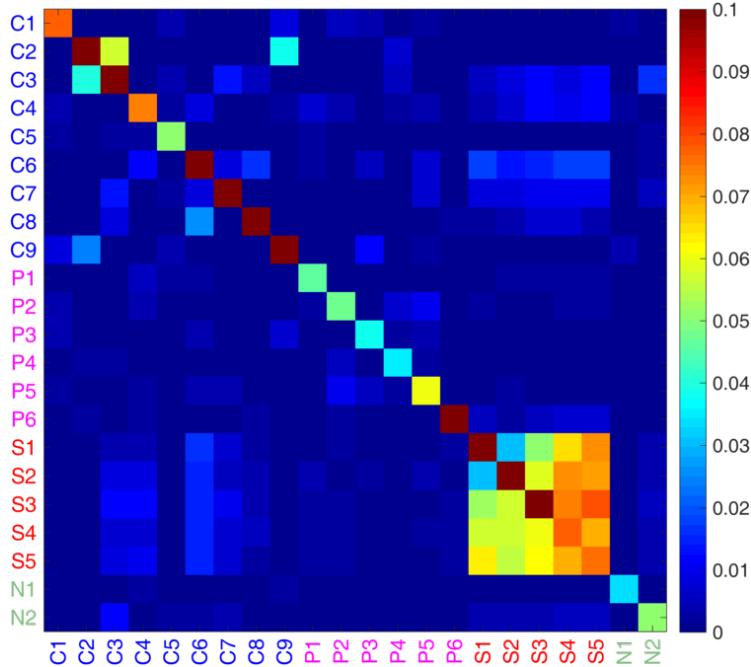


FIGURE 7.18 – Matrice de corrélation moyenne entre empreintes et résidus sur 22 sources (20 GANs et 2 caméras) [17].

L’attribution automatique par corrélation atteint une précision supérieure à 90 % même après compression JPEG (QF = 95), comme le montre la matrice de confusion Figure 7.19.

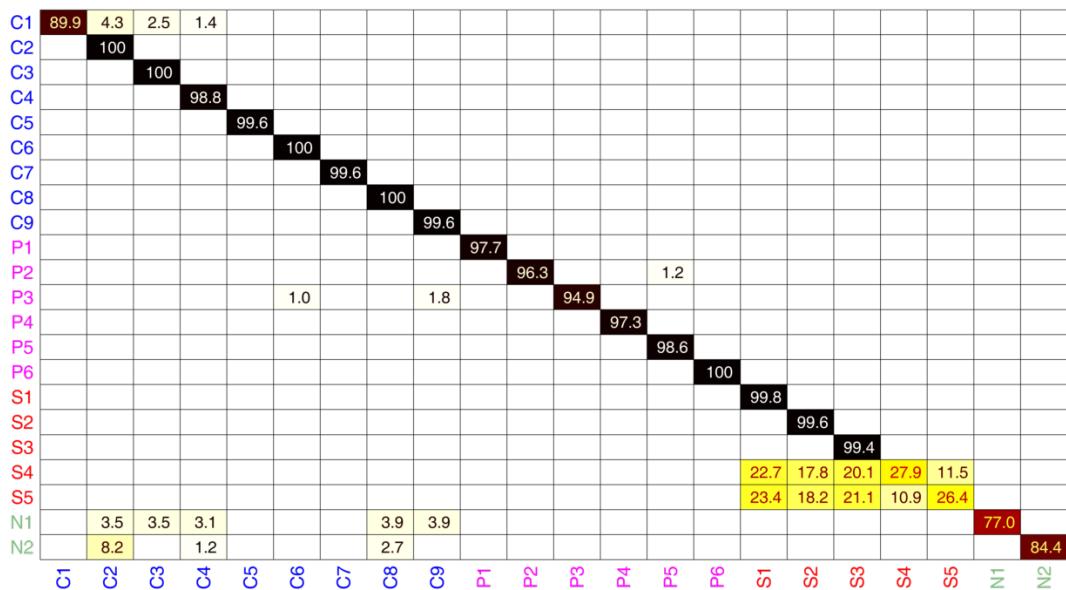


FIGURE 7.19 – Matrice de confusion pour l’attribution de source (GANs et caméras) [17].

Cette approche est appliquée avec succès au Forensics GAN Challenge (2018), où elle permet de regrouper automatiquement des images par origine GAN via finger-printing, améliorant les résultats d'un classifieur CNN via fusion (Figure 7.20).

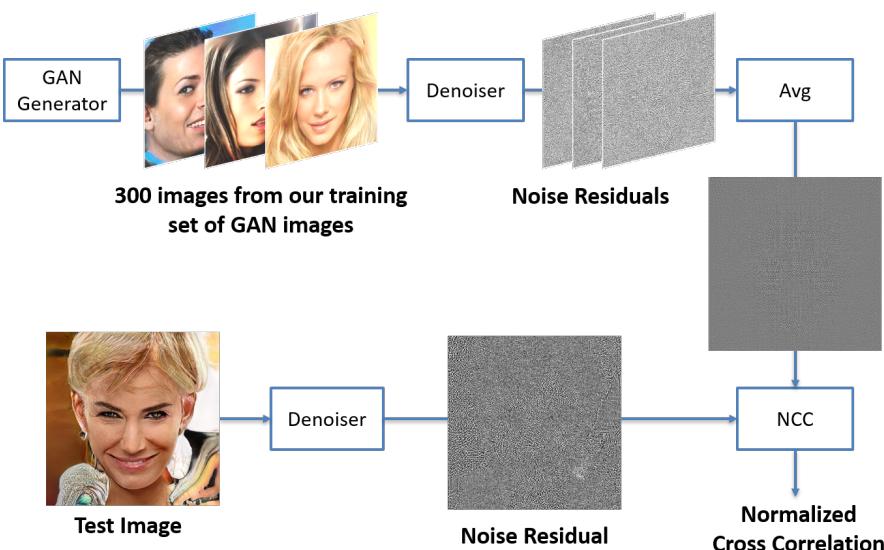


FIGURE 7.20 – Utilisation du fingerprinting dans le Forensics GAN Challenge [17].

Cette méthode se distingue par son efficacité, sa simplicité conceptuelle, et sa robustesse face à la compression. En mettant en évidence la possibilité d'extraire des empreintes spécifiques à chaque générateur, elle offre un levier puissant pour l'identification de source, la traçabilité, et la lutte contre les falsifications visuelles.

7.2.3 Analyse des points de repère faciaux

Dans une approche innovante, Li et al. proposent de détecter les visages synthétisés par GAN en s'appuyant non pas sur les artefacts visuels ou colorimétriques habituels, mais sur la configuration spatiale des points de repère faciaux (landmarks)[18]. Leur hypothèse repose sur l'observation suivante : les GAN, bien qu'excellents pour générer des détails réalistes localement, manquent souvent de contraintes globales pour positionner correctement les différentes parties du visage.

La Figure 7.21 illustre plusieurs exemples typiques d'irrégularités : asymétrie des yeux, décalage de la bouche par rapport au nez, ou angles atypiques des coins des yeux. Ces anomalies souvent subtiles à l'oeil nu, peuvent être quantifiées par l'analyse statistique des coordonnées normalisées des landmarks détectés automatiquement.



FIGURE 7.21 – Exemples d'anomalies dans des visages synthétisés par PGGAN : asymétrie des yeux (a), décalage de la bouche (b), coin de l'oeil anormalement anguleux (c) [18].

Pour détecter ces incohérences, les auteurs extraient 68 points de repère sur chaque visage, les normalisent dans un repère fixe $[0, 1] \times [0, 1]$ par transformation affine et vectorisent leurs coordonnées en un vecteur de dimension 136. Ce vecteur est ensuite utilisé comme entrée d'un classifieur SVM à noyau Radial Basis Function (RBF). Le pipeline est résumé en Figure 7.22.

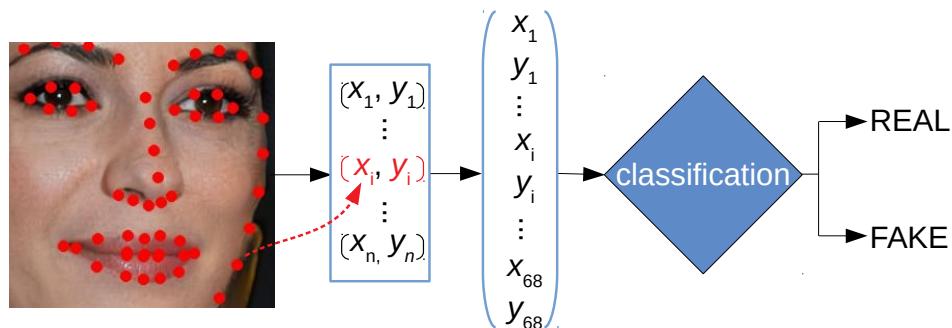


FIGURE 7.22 – Pipeline proposé : détection des landmarks, normalisation, puis classification via SVM [18].

L'analyse statistique des distributions (Figure 7.23) révèle des différences significatives entre les configurations des visages réels (CelebA) et synthétiques (PGGAN), notamment dans la distribution marginale et jointe des coordonnées.

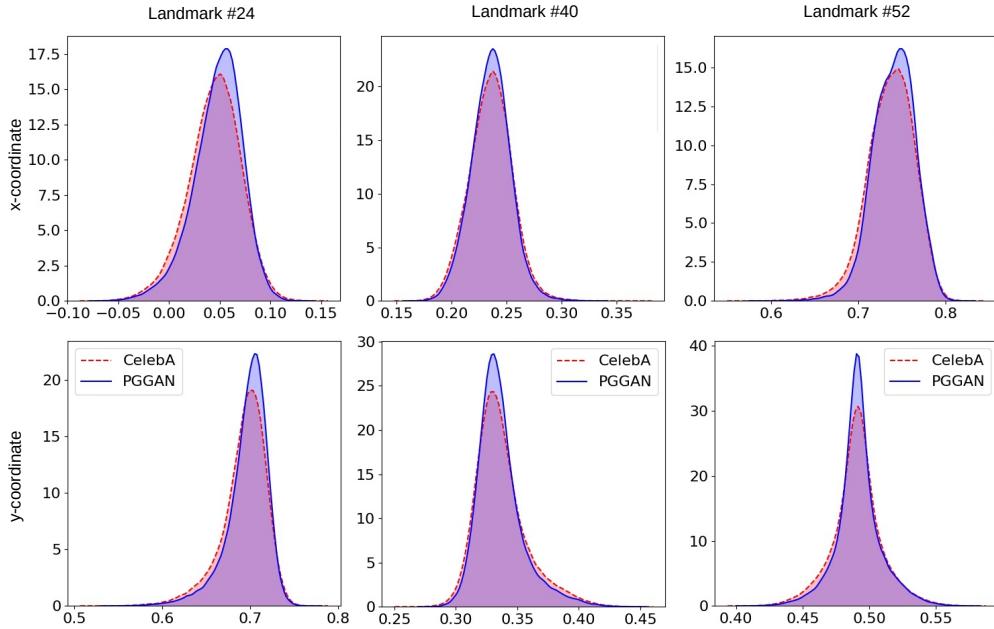


FIGURE 7.23 – Distributions des coordonnées normalisées des points de repère sur les jeux de données CelebA (réel) et PGGAN (faux) [18].

Les performances de cette méthode sont impressionnantes, d'autant plus qu'elle repose sur un modèle léger (environ 110k paramètres). Comme le montre le Tableau 7.4, le classifieur SVM atteint une *AUROC* de 94.13%, rivalisant avec des modèles profonds de plusieurs millions de paramètres.

Méthode	Paramètres	AUROC (%)
VGG19	~144M	60.13
XceptionNet	~23M	85.03
NASNet	~3.3M	96.55
ShallowNetV3	–	99.99
SVM (landmarks)	~110k	94.13

TABLE 7.4 – Comparaison des performances entre CNN et SVM sur PGGAN/CelebA [18].

Enfin les auteurs démontrent la robustesse de la méthode face aux variations de taille d'image (Figure 7.24) et montrent son application possible sur d'autres types de synthèses, notamment les vidéos Face2Face du dataset FaceForensics.

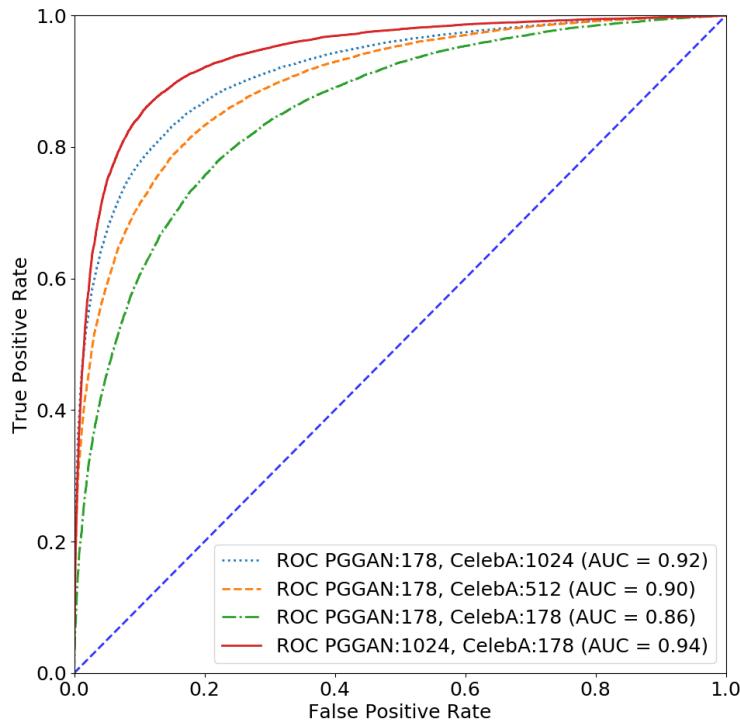


FIGURE 7.24 – Impact de la taille des images sur l’AUROC obtenu avec l’approche SVM [18].

Cette approche démontre qu’en exploitant des contraintes structurelles propres au visage humain, il est possible de détecter des visages générés artificiellement sans recourir à des réseaux profonds lourds. Elle constitue une piste précieuse pour la détection efficace et peu coûteuse de deepfakes faciaux.

7.3 Amélioration de la robustesse et généralisation

7.3.1 Détection multi-indices

La détection de fausses images repose souvent sur un seul type d’indice, qu’il soit statistique, spectral ou appris. Ce qui limite la capacité des détecteurs à généraliser face à des méthodes de génération diverses. Pour surmonter cette faiblesse, Guarnera et al. [19] proposent une approche hiérarchique fondée sur l’exploitation combinée de plusieurs modalités d’analyse, appelées multi-indices. Leur méthode structure la détection en trois niveaux successifs, allant de la distinction réel/faux jusqu’à l’identification de l’architecture de génération employée.

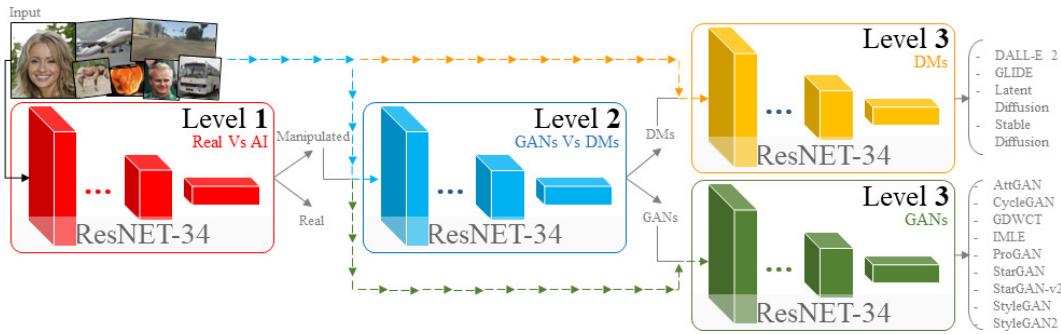


FIGURE 7.25 – Architecture d'exécution du modèle hiérarchique proposé par Guarnera et al. [19]. Le processus de classification s'effectue par niveaux : réel vs IA, puis GAN vs modèle de diffusion, et enfin identification de l'architecture générative.

Niveau 1 – Détection réel vs synthétique. À ce niveau, les images sont analysées via trois types d'indices :

- Caractéristiques colorimétriques : histogrammes RGB, HSV, YCbCr.
- Informations fréquentielles : spectre de Fourier des images.
- Descripteurs profonds : activations intermédiaires d'EfficientNet-B5.

Ces représentations sont fusionnées dans un classifieur dense pour décider si une image est réelle ou générée par une IA.

Niveau 2 – GAN vs Modèle de diffusion. Si l'image est classée comme artificielle, le système cherche à déterminer si elle a été produite par un GAN ou par un modèle de diffusion. Des caractéristiques complémentaires sont extraites pour cette tâche, avec une architecture dédiée.

Niveau 3 – Identification du générateur. Puis selon la nature du modèle identifié (GAN ou DM), un module de reconnaissance spécifique permet de discriminer entre différentes architectures (StyleGAN2, ProGAN, BigGAN, etc., ou Stable Diffusion, DALL · E, etc.).

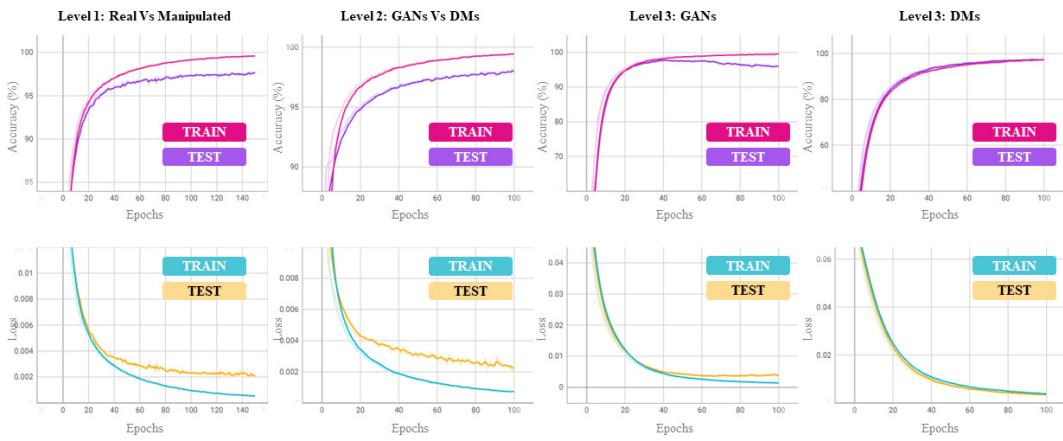


FIGURE 7.26 – Évolution de la précision et de l’erreur pour chaque niveau de classification du pipeline hiérarchique, durant l’apprentissage et le test [19].

L’approche atteint une précision moyenne de plus de 97% à chaque niveau. Elle montre notamment une excellente capacité à reconnaître les architectures GAN récentes, avec une robustesse accrue face aux stratégies de contournement.

Cette méthodologie met en lumière l’intérêt des approches multi-indices hiérarchisées : en combinant des signaux très variés, elle renforce la généralisation, réduit la dépendance à un type de trace unique et favorise une meilleure explicabilité par niveaux de décision.

Robustesse à la compression sur les réseaux sociaux

Dans une perspective complémentaire, Marra et al. [20] explorent les effets des transformations appliquées par les plateformes de réseaux sociaux (compression JPEG, redimensionnement, altération colorimétrique) sur les performances de détection. Ils simulent la diffusion d’images synthétiques et réelles sur Facebook et Twitter (X), puis évaluent différents détecteurs avant et après transmission.

Leur étude montre que les dégradations induites par ces plateformes réduisent fortement l’efficacité des détecteurs classiques, en particulier ceux reposant sur des résidus bruités ou des caractéristiques fines. Pour pallier cela, ils proposent une architecture CNN adaptée, entraînée directement sur des images post-compression.

Les résultats mettent en évidence qu’un entraînement sur des images compressées améliore nettement la robustesse des détecteurs. Ils soulignent également que certaines plateformes (comme Twitter/X) altèrent les images de manière plus agressive, ce qui nécessite des modèles spécifiquement adaptés au canal de diffusion. Cette

contribution ouvre la voie à des détecteurs *context-aware*, sensibles aux transformations subies par les images dans des scénarios réels de propagation.

7.3.2 Évaluation sur des images de haute qualité

L'émergence de GAN capables de produire des visages photoréalistes de très haute qualité, tels que StyleGAN2 ou StyleGAN3, a complexifié la tâche des détecteurs. Nowroozi et al. [21] proposent une méthode rigoureuse pour différencier les images réelles des images synthétiques générées par StyleGAN2, y compris après post-traitements. Leur étude se focalise sur des images de résolution 1024×1024 issues du dataset FFHQ, jugées visuellement indiscernables pour l'œil humain.

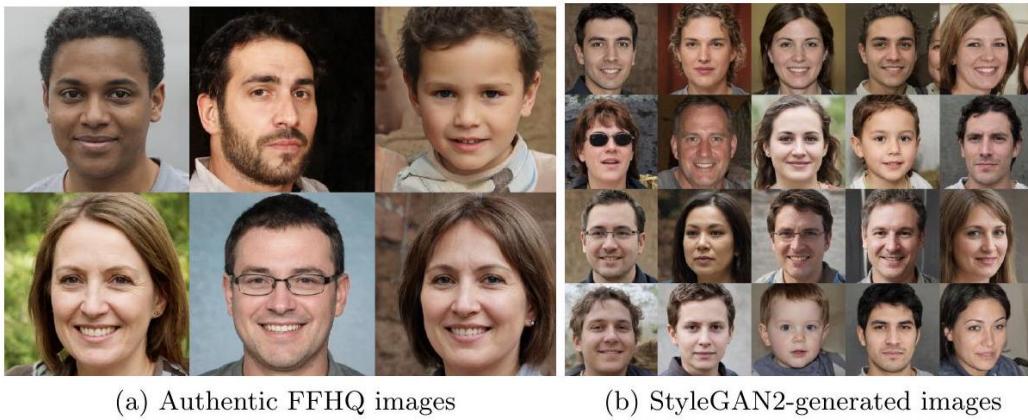


FIGURE 7.27 – Exemples de visages générés par StyleGAN2, tous de résolution 1024×1024 . Leur réalisme pose un défi majeur pour la détection automatique [21].

Les auteurs introduisent un détecteur fondé sur l'analyse conjointe des cooccurrences spatiales (intra-bandes) et croisées (inter-bandes) entre les canaux couleur. Ces matrices sont extraites de chaque image et injectées dans un réseau de neurones convolutif baptisé **Cross-Co-Net**, entraîné à partir d'une représentation en six canaux (trois intra et trois inter-bandes).

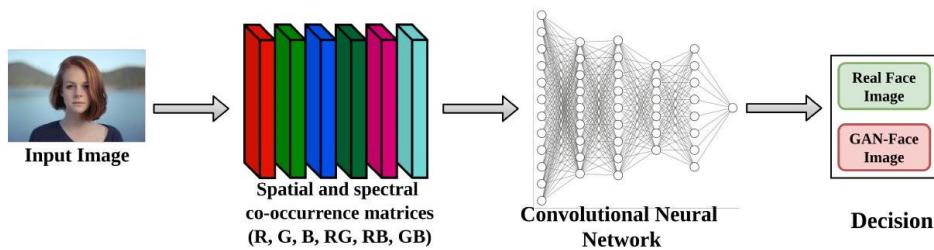


FIGURE 7.28 – Architecture du détecteur CNN proposé par Nowroozi et al., basé sur les cooccurrences spatiales et croisées entre canaux (modèle Cross-Co-Net) [21].

Les performances du modèle sont évaluées sur deux datasets complexes (StyleGAN2 et VIPPrint), et comparées à celles d'un modèle équivalent appelé Co-Net, n'utilisant que les cooccurrences intra-bandes. Sur le dataset StyleGAN2, Cross-Co-Net atteint une précision de 99,80% en détection, contre 98,25% pour Co-Net. En conditions JPEG-aware (entraînement sur images compressées), Cross-Co-Net maintient une précision moyenne de 94,40%.

Les expériences montrent également une robustesse supérieure de Cross-Co-Net face à divers post-traitements : flou, bruit, redimensionnement, rotation, histogram equalization. Cette stabilité est particulièrement notable là où les méthodes traditionnelles échouent (résultats proches de 50% pour Co-Net dans les cas extrêmes).

Ce travail met en lumière l'intérêt d'exploiter des cooccurrences croisées entre canaux pour capter des irrégularités subtiles dans les images GAN, tout en renforçant la robustesse aux dégradations typiques rencontrées lors de la diffusion des images. Il constitue une piste prometteuse pour les détecteurs futurs, notamment dans un contexte de haute résolution et de compression.

7.4 Enjeux de résistance aux contre-attaques

Les méthodes de détection basées sur l'apprentissage supervisé restent vulnérables à divers types de contre-attaques, qu'elles soient accidentelles (compression, redimensionnement, bruit) ou intentionnelles (perturbations adversariales). Ces transformations visent à altérer les caractéristiques exploitées par les détecteurs sans affecter la perceptibilité humaine, compromettant ainsi leur efficacité dans des conditions réelles.

7.4.1 Effets des post-traitements standards

Plusieurs travaux ont montré que des traitements classiques comme la compression JPEG, le recalibrage de la luminosité, le flou gaussien ou l'ajout de bruit pouvaient affecter la performance des détecteurs de fausses images. Par exemple, Marra et al. [20] simulent la diffusion d'images sur les réseaux sociaux en les faisant transiter par les plateformes Facebook et Twitter (X). Ils observent que les détecteurs voient leur précision chuter significativement après ces traitements, en particulier à cause de la compression agressive appliquée côté serveur.

Nowroozi et al. [21] confirment cette tendance sur des images haute résolution générées par StyleGAN2. Ils proposent un détecteur appelé Cross-Co-Net, basé sur l'analyse des cooccurrences intra- et inter-bandes, et le comparent à un modèle plus simple, Co-Net. Les deux réseaux sont soumis à des perturbations progressives telles

que le bruit gaussien, le flou, le redimensionnement, la rotation ou la compression JPEG.

Les résultats expérimentaux montrent que Co-Net s'effondre rapidement (jusqu'à 50% de précision) dès qu'un post-traitement modifie les artefacts exploités. En revanche, Cross-Co-Net conserve une précision largement supérieure dans tous les cas, dépassant 90% dans la majorité des scénarios. L'étude souligne notamment que les relations croisées entre canaux sont moins sensibles aux dégradations purement spatiales, ce qui confère à Cross-Co-Net une meilleure résilience aux altérations visuelles classiques.

Ces résultats plaident pour une meilleure intégration de la diversité des transformations possibles dans les protocoles d'entraînement, ainsi que pour des architectures capables de s'appuyer sur des indices plus globaux et plus robustes aux perturbations.

TABLE 7.5 – Robustesse comparée des modèles Cross-Co-Net et Co-Net sous divers post-traitements (StyleGAN2).

Post-traitement	Cross-Co-Net (%)	Co-Net (%)
Flou (3×3)	92.85	72.50
Bruit gaussien ($\sigma = 2$)	90.70	50.00
Compression JPEG (QF = 85)	96.28	93.08
Redimensionnement ($\times 0.5$)	81.50	50.05
Égalisation d'histogramme (AHE)	75.00	50.00
Rotation (45°)	99.50	71.90

7.4.2 Perturbations adversariales ciblées

Au-delà des dégradations involontaires, des attaques adversariales peuvent être spécifiquement conçues pour tromper les détecteurs. Dans ce cas l'image est légèrement modifiée de manière imperceptible pour l'humain, mais suffisante pour induire une mauvaise classification du modèle.

Bien que peu d'études se concentrent directement sur l'adversarialité dans le contexte des GAN, plusieurs publications sur la détection de manipulations numériques [41] ou sur la stéganalyse [38, 31] ont montré la sensibilité des détecteurs à des modifications de type gradient-based. Cela soulève la question de la robustesse des détecteurs CNN aux signaux non attendus ou adversariaux.

Certaines contre-mesures proposées incluent :

- l’entraînement avec data augmentation extensive, incluant des images compressées, bruitées, retournées ou redimensionnées,
- l’utilisation de modèles JPEG-aware [30], sensibles à la structure même de la compression,
- la fusion de modalités diverses, comme dans les approches multi-indices de Guarnera et al. [19], afin de renforcer la généralisation par redondance.

7.4.3 Limites et perspectives

La combinaison de ces travaux révèle un point critique : un détecteur efficace en laboratoire peut devenir inopérant dans un contexte réaliste. Il devient donc essentiel d’intégrer dans la phase d’entraînement une variété de perturbations, de simuler les parcours de diffusion réels (réseaux sociaux, compression mobile), et d’explorer des méthodes de normalisation ou de défense adversariale.

Plusieurs pistes sont actuellement explorées :

- la détection robuste aux conditions de transmission (recherche d’invariances visuelles) ;
- les architectures hybrides (combinaison de CNN fixes et adaptatifs) ;
- les méthodes de détection adversariale défensives (adversarial training, distillation, etc.).

Dans ce contexte, la détection de fausses images devient un problème à la fois dynamique, contextuel, et fondamentalement lié aux conditions de traitement des images.

Chapitre 8

Conclusion

À l'heure où les modèles génératifs atteignent un niveau de réalisme inédit, la distinction entre le vrai et le faux visuel devient chaque jour plus difficile à établir. Ce mémoire s'est attaché à explorer différentes méthodes de détection d'images générées par des réseaux antagonistes génératifs (GAN), en mettant en lumière à la fois les progrès réalisés et les nombreuses limites encore présentes.

Nous avons d'abord retracé les grandes étapes de la falsification visuelle avant l'émergence des GAN, montrant que si les manipulations d'images ne sont pas nouvelles, la facilité et l'efficacité offertes par les GAN ont radicalement changé la donne. Ensuite, nous avons détaillé les fondements des modèles génératifs, jusqu'à l'apparition des GAN et de leurs nombreuses variantes, notamment les architectures modernes telles que StyleGAN, qui repoussent toujours plus loin les frontières du réalisme.

Ce réalisme accru a logiquement suscité l'émergence de multiples méthodes de détection, que nous avons classées en trois grandes familles : les approches sans apprentissage profond (fondées sur des artefacts statistiques ou des propriétés physiques), les méthodes utilisant l'apprentissage automatique (notamment les réseaux convolutifs ou l'analyse des empreintes GAN), et enfin, les approches cherchant à renforcer la robustesse des détecteurs face aux attaques, à la compression ou à la généralisation inter-modèles.

L'ensemble de ces techniques témoigne de la complexité croissante du problème :

chaque amélioration du réalisme visuel apportée par les GAN engendre une réponse adaptative de la part des chercheurs en forensique numérique. Pourtant, comme nous l'avons vu, aucune méthode ne se révèle infaillible, et de nombreux défis restent à relever. En particulier, la question de la généralisation, c'est-à-dire la capacité d'un détecteur à fonctionner sur des images issues de modèles inconnus demeure un point faible crucial. De même, la résilience des détecteurs aux traitements postérieurs tels que la compression ou la redimension d'image constitue une difficulté technique majeure.

Par ailleurs, nos analyses sur la perception humaine ont souligné une réalité inquiétante : l'humain seul est souvent incapable de discerner une image authentique d'une image générée, même lorsqu'il dispose de temps ou d'indices contextuels. Ce constat vient renforcer l'idée que des outils techniques sont désormais indispensables pour accompagner les utilisateurs, professionnels ou grand public, dans l'identification des contenus visuels falsifiés.

Ce travail ouvre ainsi plusieurs perspectives. D'une part, l'exploration de méthodes hybrides combinant indices physiques, réseaux de neurones, et analyse perceptive pourrait offrir des pistes prometteuses. D'autre part, l'intégration de systèmes de détection directement dans les plateformes de diffusion (réseaux sociaux, moteurs de recherche) représente un enjeu majeur de transfert technologique et de gouvernance.

Enfin, au-delà des aspects techniques, ce mémoire invite à une réflexion éthique plus large. La prolifération d'images synthétiques interroge notre rapport à la vérité, à la preuve visuelle, et plus largement à la confiance numérique. La lutte contre les fausses images générées par GAN ne pourra donc être menée efficacement sans une collaboration étroite entre chercheurs, ingénieurs, juristes, et acteurs de la société civile.

Références

- [1] V. Schetinger, M. M. Oliveira, R. da Silva, and T. J. Carvalho, “Humans are easily fooled by digital images,” 2015. [Cité en pages ix, xv, 3, 4, 5 et 6]
- [2] T. Neubert, “Face morphing detection : An approach based on image degradation analysis,” pp. 93–106, 07 2017. [Cité en pages ix, 2, 12, 13, 14 et 15]
- [3] R. Wu, X. Li, and B. Yang, “Identifying computer generated graphics via histogram features,” in *2011 18th IEEE International Conference on Image Processing*, pp. 1933–1936, 2011. [Cité en pages ix, 10, 17, 18 et 19]
- [4] J. Fridrich, M. Goljan, and D. Hogea, “Steganalysis of jpeg images : Breaking the f5 algorithm,” in *International Workshop on Information Hiding*, pp. 310–323, Springer, 2002. [Cité en pages ix, 21 et 22]
- [5] S. Bayram and N. Memon, “An efficient and robust method for detecting copy-move forgery,” in *IEEE Transactions on Information Forensics and Security*, vol. 3, pp. 65–74, IEEE, 2009. [Cité en pages ix et 24]
- [6] L. A. Gatys, A. S. Ecker, and M. Bethge, “A neural algorithm of artistic style,” 2015. [Cité en pages ix, 28, 32 et 33]
- [7] M. Probst and F. Rothlauf, “Training a restricted boltzmann machine for classification by labeling model samples,” 2015. [Cité en pages ix, 34 et 35]
- [8] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial networks,” 2014. [Cité en pages x, 1, 7, 37, 39, 40 et 41]
- [9] A. Radford, L. Metz, and S. Chintala, “Unsupervised representation learning with deep convolutional generative adversarial networks,” 2016. [Cité en pages x, 42, 43, 44, 45 et 46]
- [10] T. Karras, T. Aila, S. Laine, and J. Lehtinen, “Progressive growing of gans for improved quality, stability, and variation,” 2018. [Cité en pages x, 46, 47, 48 et 60]
- [11] T. Karras, S. Laine, and T. Aila, “A style-based generator architecture for generative adversarial networks,” 2019. [Cité en pages x, 49, 50, 51 et 52]
- [12] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, “Analyzing and improving the image quality of stylegan,” 2020. [Cité en pages xi, 51 et 52]
- [13] L. Nataraj, T. M. Mohammed, S. Chandrasekaran, A. Flenner, J. H. Bappy, A. K. Roy-Chowdhury, and B. S. Manjunath, “Detecting gan generated fake images using co-occurrence matrices,” 2019. [Cité en pages xi, 2, 55, 56 et 57]

- [14] S. McCloskey and M. Albright, “Detecting gan-generated imagery using color cues,” 2018. [Cité en pages xi, 2, 58, 59 et 60]
- [15] X. Zhang, S. Karaman, and S.-F. Chang, “Detecting and simulating artifacts in gan fake images,” 2019. [Cité en pages xi, xii, 62, 63, 64, 65 et 66]
- [16] D. Cozzolino, G. Poggi, and L. Verdoliva, “Recasting residual-based local descriptors as convolutional neural networks : an application to image forgery detection,” 2017. [Cité en pages xii, 67, 68 et 69]
- [17] F. Marra, D. Gragnaniello, L. Verdoliva, and G. Poggi, “Do gans leave artificial fingerprints ?,” 2018. [Cité en pages xii, 2, 70, 71, 72 et 73]
- [18] X. Yang, Y. Li, H. Qi, and S. Lyu, “Exposing gan-synthesized faces using landmark locations,” 2019. [Cité en pages xii, xv, 74, 75 et 76]
- [19] L. Guarnera, O. Giudice, and S. Battiato, “Level up the deepfake detection : a method to effectively discriminate images generated by gan architectures and diffusion models,” 2023. [Cité en pages xii, 76, 77, 78 et 82]
- [20] F. Marra, D. Gragnaniello, D. Cozzolino, and L. Verdoliva, “Detection of gan-generated fake images over social networks,” pp. 384–389, 04 2018. [Cité en pages 78 et 80]
- [21] E. Nowroozi, M. Conti, and Y. Mekdad, “Detecting high-quality gan-generated face images using neural networks,” 2022. [Cité en pages xii, xiii, 79 et 80]
- [22] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, “Deepfakes and beyond : A survey of face manipulation and fake detection,” 2020. [Cité en page 1]
- [23] G. Wolberg, *Digital Image Warping*. Los Alamitos, CA, USA : IEEE Computer Society Press, 1990. [Cité en page 2]
- [24] G. Wolberg, “Image morphing : A survey,” *The Visual Computer*, vol. 14, 03 1999. [Cité en pages 2, 12 et 13]
- [25] A. Lamb, “A brief introduction to generative models,” 2021. [Cité en pages 2, 27, 28, 31, 32, 34 et 35]
- [26] C. Renouard, “L’image composite dans les effets spéciaux visuels cinématographiques, de l’analogue au numérique,” *Sens public*, p. Articles 1582, Jan. 2022. [Cité en page 10]
- [27] D. Zlatkova, P. Nakov, and I. Koychev, “Fact-checking meets fauxtography : Verifying claims about images,” 2019. [Cité en pages 11 et 15]
- [28] ITU-T, “Digital compression and coding of continuous-tone still images—requirements and guidelines.” <https://www.w3.org/Graphics/JPEG/itu-t81.pdf>, 1992. ITU-T Recommendation T.81, also ISO/IEC 10918-1 :1993. [Cité en page 20]

- [29] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*. Prentice Hall, 2nd ed., 2002. [Cité en page 20]
- [30] J. Fridrich, M. Goljan, and R. Du, "Steganalysis based on jpeg compatibility," in *Proceedings of SPIE - The International Society for Optical Engineering*, vol. 4518, pp. 275–280, International Society for Optics and Photonics, 2001. [Cité en pages 21 et 82]
- [31] T. Pevný, P. Bas, and J. Fridrich, "Steganalysis by subtractive pixel adjacency matrix," *IEEE Transactions on Information Forensics and Security*, vol. 5, no. 2, pp. 215–224, 2010. [Cité en pages 23 et 81]
- [32] J. Lukáš, J. Fridrich, and M. Goljan, "Digital camera identification from sensor pattern noise," in *IEEE Transactions on Information Forensics and Security*, vol. 1, pp. 205–214, IEEE, 2006. [Cité en page 23]
- [33] A. Elgammal, B. Liu, M. Elhoseiny, and M. Mazzone, "Can : Creative adversarial networks, generating "art" by learning about styles and deviating from style norms," 2017. [Cité en page 28]
- [34] A. van den Oord, N. Kalchbrenner, O. Vinyals, L. Espeholt, A. Graves, and K. Kavukcuoglu, "Conditional image generation with pixelcnn decoders," 2016. [Cité en page 34]
- [35] W. Fedus, M. Rosca, B. Lakshminarayanan, A. M. Dai, S. Mohamed, and I. Goodfellow, "Many paths to equilibrium : Gans do not need to decrease a divergence at every step," 2018. [Cité en page 41]
- [36] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," 2017. [Cité en page 49]
- [37] A. Sullivan, U. Madhow, S. Chandrasekaran, and B. S. Manjunath, "Steganalysis of jpeg images using statistical and image quality features," in *IEEE International Conference on Image Processing 2005*, vol. 2, pp. II–1282, IEEE, 2005. [Cité en page 55]
- [38] J. Fridrich and J. Kodovsky, "Rich models for steganalysis of digital images," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 3, pp. 868–882, 2012. [Cité en pages 55 et 81]
- [39] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2223–2232, 2017. [Cité en page 57]
- [40] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "Stargan : Unified generative adversarial networks for multi-domain image-to-image translation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8789–8797, 2018. [Cité en page 57]

- [41] D. Cozzolino, G. Poggi, and L. Verdoliva, “Recasting residual-based local descriptors as convolutional neural networks : an application to tampering localization,” *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, 2017. [Cité en pages 58 et 81]
- [42] C. Chen, S. McCloskey, and J. Yu, “Focus manipulation detection via photometric histogram analysis,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. [Cité en page 61]
- [43] L. Verdoliva, “Media forensics and deepfakes : An overview,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, p. 910–932, Aug. 2020.