Distance image-image pour la détection d'images générées par modèles de diffusion

Antoine MALLET[1] Patrick BAS[2] Rémi COGRANNE[1]

[1]Université de Technologie de Troyes, LIST3N, 12 Rue Marie Curie, CS42060, 10004 Troyes Cedex, France

[2]Univ. Lille, CNRS, Centrale Lille, UMR 9189 CRIStAL, Bât. ESPRIT, Avenue Henri Poincaré 59655 Villeneuve d'Ascq

Résumé – Ce travail propose une méthode de détection d'images générées par modèles de diffusion. Dans un premier temps, nous proposons une mesure statistique de la distribution multivariée du résidu des pixels par un filtre passe-haut. Cette mesure sert d'empreinte digitale de la nature réelle ou générée d'une image. Dans un second temps, nous rappelons comment mesurer une distance dans la variété des statistiques extraites, en l'occurence des matrices de corrélations. Puis nous présentons un schéma de détection, en mesurant la distance des statistiques extraites sur une base d'images, et sur les images obtenues après une itération de modèles de diffusion « image-to-image ». Les modèles de diffusion marquent ainsi l'empreinte de l'image, sans modifier le contenu. Nous montrons que la marque sur les images réelles est plus importante que sur les images générées. Les résultats confirment l'intérêt de l'approche dans un problème de classification binaire, et les limites en multiclasse.

Abstract – This work proposes a method for detecting images generated by diffusion models. Firstly, we propose a statistical measure of the multivariate distribution of the pixel residual using a high-pass filter. This measure serves as a fingerprint for the real or synthetic nature of an image. We then review how to measure a distance in the manifold of the extracted statistics, in this case correlation matrices. We then present a detection scheme, by measuring the distance between the extracted statistics on an image base, and on images obtained after an iteration of image-to-image diffusion models. The diffusion models thus mark the fingerprint of the image, without modifying the content. We show that the imprint on real images is greater than on synthetic images. The results confirm the interest of the approach for diffusion models, as well as the limitations for other types of generators.

1 Introduction

Les outils de génération automatique d'images par apprentissage profond offrent un acccès inédit à la production visuelle et artistique. Cette avancée technologique représente certes une opportunité, mais présente aussi un enjeu relatif à la falsification de l'information, voire à la propagation de contenus néfastes. Récemment, les modèles de diffusion (DMs) ont surclassés les réseaux adverses génératifs (GANs) par la versatilité des contenus générés, mais aussi par leur flexibilité d'utilisation, en mode « *text-to-image* » ou en mode « *image-to-image* », tout en maintenant un photoréalisme toujours croissant. Face à cette prolifération, le besoin d'identifier les contenus générés est appraru comme un enjeu sociétal important. En réponse, le milieu académique a produit ces dernières années une littérature foisonnante. Même si dans un premier temps, des indicateurs basés sur des incohérences sémantiques ont suffi [13], la qualité visuelle grandissante des images produites a rapidement déplacé leur détection vers des analyses plus fines.

Des modèles d'apprentissage profond ont été proposés pour la détection d'images de GANs, comme un ResNet-50 [15], puis pour des images de DMs [6]. Comme dans toute tâche d'apprentissage, le surapprentissage a imposé des limites à la généralisation de ces modèles. De nombreux travaux ont ainsi proposé des approches forensiques, toutefois souvent en partie construite sur de l'apprentissage profond. Dans le domaine spatial, des matrices de co-occurences sur les canaux couleurs ont été proposées [14]. Dans le domaine fréquentiel, la visualisation du sur-échantillonnages des modèles génératifs dans les spectres de Fourier est une illustration populaire [1]. Bien qu'initialement proposée pour les GANs [12], cette observation s'est maintenue pour les DMs [5], chez lesquels on peut interpréter le passage de l'espace latent vers l'espace pixel comme un sur-échantillonnage. Ainsi, [9] propose d'apprendre le module d'extraction des fréquences caractéristiques des images pour la détection des images générées. Récemment, la proposition d'utiliser la représentation des images dans des espaces latents a montré des résultats prometteurs. En utilisant CLIP, [7] montre qu'un SVM offre des bons résultats. Plus récemment encore, [3] propose un espace latent issu d'un apprentissage par contraste, maximisant la distance entre images réelles et générées. Cet écart est également exploité par [16], qui montre que le processus de diffusion modifie davantage une image réelle qu'une image générée.

Ainsi, en génération « *image-to-image* », l'impact du processus de diffusion semble être plus marqué pour les images réelles que pour les images issues de modèles de diffusion. Nous proposons dans cet article d'exploiter cette différence pour la détection d'images générées par des modèles de diffusion. Au-delà des performances, nous souhaitons mettre en avant une preuve de concept basée sur un très petit nombre de dimensions pour réaliser la détection. La méthode proposée repose sur une empreinte statistique de l'image et sur une distance entre ces statistiques. Une régression logistique est utilisée sur la distribution des distances de chaque classe pour faire de la classification binaire. En particulier, nous mettons en exergue la complémentarité des canaux couleurs.

Le reste de cet article est organisé comme suit. La section 2 détaille la méthode proposée. La section 3 présente les conditions expérimentales et les résultats obtenus. La section 4 conclut et dessine les perspectives offertes par ce travail.

2 Méthodologie

2.1 Empreinte statistique résiduelle

Les bruits des images générées et réelles diffèrent, comme souvent illustrés par les spectres de Fourier. Une approche pratique pour capturer une empreinte synthétique et discriminante est de regarder la covariance locale du bruit d'une image. Cette approche a déjà été proposée pour l'identification des « *cover-sources* » en stéganalyse [10]. Elle consiste dans un premier temps à calculer le résidu $\mathcal{F}$ d'une image $I \in \mathcal{M}^{N\times N}$, approximation du bruit, naturellement spatialement multivarié [8]. Nous proposons, pour calculer ce résidu, l'opérateur $\mathcal{L}$ suivant :

$$\mathcal{F} = I \circledast \mathcal{L} = I \circledast \begin{pmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{pmatrix}, \quad (1)$$

où $\circledast$ est l'opérateur de convolution. Cette estimation triviale est suffisante pour capturer les variations locales du bruit [10]. De ce résidu, nous construisons des échantillons de bruit $b_i^{\mathcal{F}} \in \mathbb{R}^p$, qui sont des patchs vectorisés de 8×8 pixels du résidu $\mathcal{F}$, ce qui donne $p = 64$. Nous sélectionnons ensuite parmi eux ceux portant le moins de contenu, c'est-à-dire les échantillons ayant conjointement une faible moyenne et une faible variance :

$$\mathcal{B}^{\mathcal{F}} = \left\{ b_k^{\mathcal{F}} \;/\; b_k^{\mathcal{F}} \in \min_k^{\tau} \mathbb{V}\left[b_k^{\mathcal{F}}\right] \bigcap \min_k^{\tau} \mathbb{E}\left[b_k^{\mathcal{F}}\right] \right\}, \quad (2)$$

où $\min_k^{\tau}$ renvoie les τ plus petites valeurs sur k, $\mathbb{V}[.]$ la variance et $\mathbb{E}[.]$ la moyenne. Enfin, la matrice de corrélation $\mathbf{\Sigma} = (\sigma)_{i,j}, (i,j) \in \{1, \dots 64\}^2$ du résidu est calculée sur les $n \leq \tau$ échantillons de $\mathcal{B}^{\mathcal{F}}$:

$$(\sigma)_{i,j} = \frac{\sum_{k=1}^{n} (b_k^{\mathcal{F}} - \overline{b_k^{\mathcal{F}}})(b_k^{\mathcal{F}} - \overline{b_k^{\mathcal{F}}})^T}{\sqrt{\sum_{k=1}^{n} (b_{k,i}^{\mathcal{F}} - \overline{b_{k,i}^{\mathcal{F}}})^2 \sum_{k=1}^{n} (b_{k,j}^{\mathcal{F}} - \overline{b_{k,j}^{\mathcal{F}}})^2}}, \quad (3)$$

où $\overline{b_k^{\mathcal{F}}}$ est la moyenne de $b_k^{\mathcal{F}}$. Des exemples de matrices obtenues sont présentés en figure 1. Pour simplifier, comme les structures se répètent le long de la diagonale, nous n'en présentons que le quart supérieur gauche. Une cellule rouge indique une corrélation positive entre deux variables, une cellule bleue une corrélation négative. Des études ont montré que ces coefficients discriminent les images générées des images réelles [11].

2.2 Distance entre matrices de corrélations

Une matrice de corrélation est une matrice symétrique définie semi-positive. Elle vit donc dans une variété $\mathcal{P}$ de dimension $\frac{p(p+1)}{2}$. Sur cette variété, la distance entre deux points $\mathbf{\Sigma_1}$ et $\mathbf{\Sigma_2}$ est donnée par la distance géodésique $\delta_{\mathcal{P}}$ suivante [2] :

$$\delta_{\mathcal{P}}(\mathbf{\Sigma_1}, \mathbf{\Sigma_2}) = \| \log(\mathbf{\Sigma_1^{-1}\Sigma_2}) \|_F = \left[\sum_{i=1}^{k} \log^2 \lambda_i \right]^{\frac{1}{2}}, \quad (4)$$

où $\|.\|_F$ est la norme de Frobenius et $\lambda_i, i = 1, \dots, k$ sont les valeurs propres de $\mathbf{\Sigma_1^{-1}\Sigma_2}$.

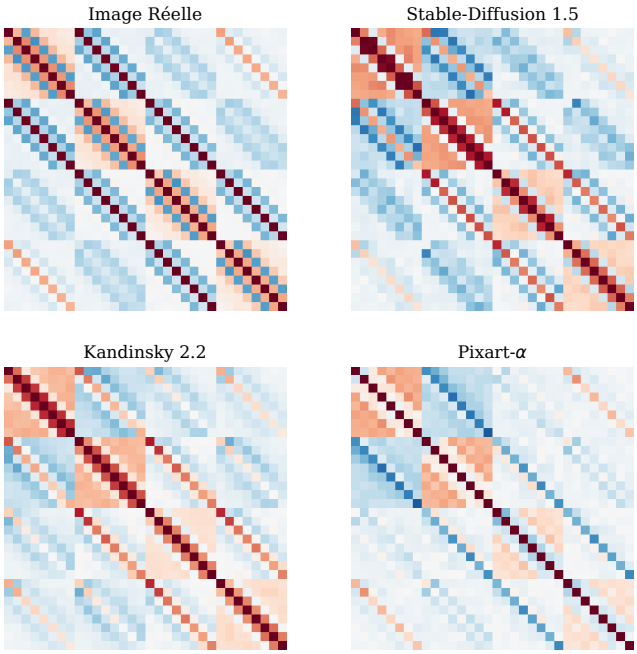

FIGURE 1 : Exemples de matrices de corrélation obtenues sur des images réelles et générées. La statistique du canal luminance est présentée.

2.3 Modèle de détection

Soient les corrélations $\Sigma_{(r)}$ et $\Sigma_{(f)}$ estimées sur les résidus d'une image réelle et d'une image générée, respectivement. Soient $\Sigma'_{(r)}$ et $\Sigma'_{(f)}$ les corrélations estimées sur les résidus des images obtenues après une itération d'un modèle de diffusion appliquée aux deux images précédentes. Dans ce travail, nous comparons les distances géodésiques $\delta_{\mathcal{P}}(\Sigma_{(r)}, \Sigma'_{(r)})$ et $\delta_{\mathcal{P}}(\Sigma_{(f)}, \Sigma'_{(f)})$.

Notons les hypothèses $\mathcal{H}_0$: « l'image initiale est réelle », et $\mathcal{H}_1$: « l'image initiale est générée ». Pour discriminer un échantillon entre ces deux hypothèses, nous utiliserons une régression logistique apprise sur une base d'entraînement de distances géodésiques, et choisirons le seuil minimisant la probabilité d'erreur totale sous hypothèse d'équiprobabilité des classes.

La figure 2 illustre la distance géodésique entre une paire « réelle-diffusée » et une paire « générée-diffusée ». Suivant les observations de [16], nous pouvons schématiquement représenter les éléments d'une paire « générée-diffusée » comme plus proches que ceux d'une paire « réelle-diffusée ».

3 Résultats expérimentaux

3.1 Cadre expérimental

Notre expérience est réalisée sur une base d'images couleurs de dimension 512×512 au format PNG. La classe $\mathcal{H}_0$ est consituée de 10.000 images réelles issues de la base d'image ALASKA [4]. La classe $\mathcal{H}_1$ est constituée d'images générées issues des modèles de diffusion suivants : Kandinsky 2.2, DreamLike-Photoreal 2.0, Pixart-α, Stable-Diffusion 1.5 et Stable-Diffusion 2.1. Pour chacun de ces modèles, nous avons généré 2.000 images, soit 10.000 au total.

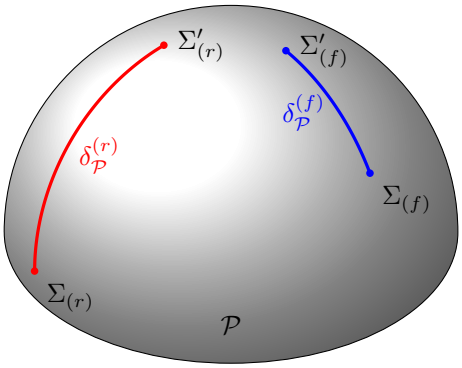

FIGURE 2 : Représentation schématique de la distance géodésique entre deux matrices de corrélation, représentée par les arcs de cercles $\delta_{\mathcal{P}}^{(r)}$ et $\delta_{\mathcal{P}}^{(f)}$. Les points de $\mathbb{R}^3$ vivent dans la variété $\mathcal{P} \subset \mathbb{R}^2$. Ici, $\Sigma_{(r)}$ est issue d'une image réelle, $\Sigma'_{(r)}$ de l'image diffusée à partir de l'image réelle.

Pour générer la base de distances, nous utilisons trois modèles de diffusion en mode « *image-to-image* » : Kandinsky 2.2, Stable-Diffusion 1.5 et DreamLike-Photoreal 2.0. Nous obtenons ainsi, pour chacune des 20.000 images initiales, 3 images diffusées avec une seule itération.

3.2 Analyse des canaux couleurs

Dans un premier temps, nous proposons d'étudier la distribution des distances obtenues sur chaque canal couleur, et avec chaque générateur. Les histogrammes sont donnés en figure 3. Nous voyons que la distribution des distances diffère d'un modèle *image-to-image* à l'autre. En particulier, elle semble moins discriminantes pour le modèle Kandinsky 2.2 (milieu), où la queue droite de la distribution des images générées est beaucoup plus épaisse, que pour les modèles DreamLike-Photoreal 2.0 (haut) et Stable-Diffusion 1.5 (bas). Ces queues de distributions sont expliquées par l'inconsistance des distances issues des modèles « *image-to-image* » selon le générateur utilisé pour les images initiales. Cette inconsistance est illustrée par la figure 4. Les histogrammes de la luminance (colonne de gauche de la figure 3) y sont détaillés pour chaque générateur. On y voit effectivement que pour le modèle Kandinsky 2.2, la distribution des distances pour les images initiales de Stable-Diffusion 2.1 et Pixart-α est très peu discernable de celle des images réelles. A l'inverse, on observe que les images initialement issues de DreamLike-Photoreal 2.0 et Stable-Diffusion 1.5 sont relativement bien discernables des images réelles, peu importe que les distances soient calculées à travers l'un ou l'autre des modèles « *image-to-image* ». Cela s'explique par le fait que les deux architectures de diffusion sont en réalité similaires.

Pour quantifier la détectabilité entre $\mathcal{H}_0$ et $\mathcal{H}_1$ à partir de ces distances, nous entraînons une régression logistique sur chacun des 9 canaux séparément, puis par modèle « *image-to-image* » (Y-Cr-Cb agrégés), puis par canal couleur (Dreamlike-Kandinsky-Stable-Diffusion agrégés), puis tous ensemble. Les résultats sont donnés en table 1. Nous obs ervons que les chrominances sont de manière générale plus discriminantes que la luminance, qu'elles soient prises individuellement (DreamLike et Kandinsky) ou agrégées. Comme observé sur les histogrammes, le modèle Kandinsky 2.2 est moins

Canal	Dreamlike	Kandinsky	Stable-Diff	Tous
Y	25.8	48.0	25.9	20.8
Cr	18.1	35.3	25.8	17.9
Cb	18.9	37.1	28.4	17.2
Tous	14.1	28.6	21.4	11.7

TABLE 1 : Probabilité d'erreur (en %) de la classification binaire pour chaque canal couleur et chaque générateur en entrée. Une validation croisée à 10 plis est utilisée.

discriminant que les autres, et ce sur tous les canaux couleurs. Enfin, l'agrégation de tous les canaux couleurs donne les meilleurs résultats, avec une probabilité d'erreur moyenne de 11.7%. Il s'agit donc d'une performance relativement bonne, compte tenu de la simplicité de la méthode, et du petit nombre de dimensions utilisées.

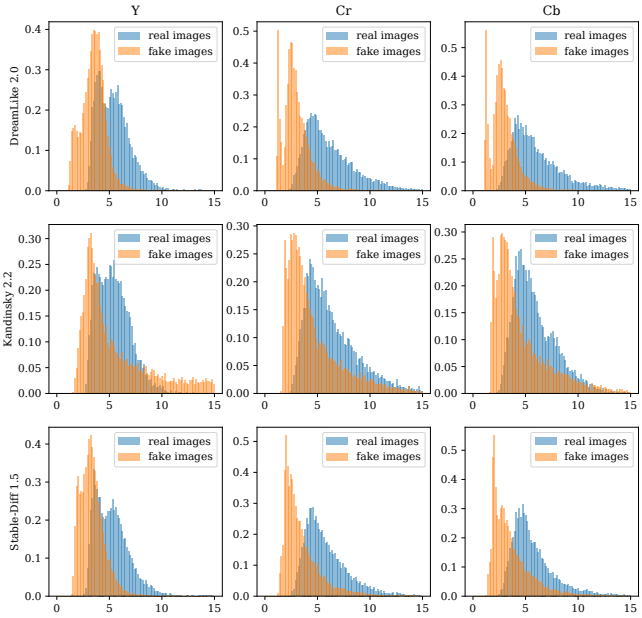

FIGURE 3 : Histogrammes des distances géodésiques obtenues sur les canaux couleurs pour chaque générateur. En bleu, les distances pour les paires « réelle-diffusée » ($\mathcal{H}_0$), en orange les distances pour les paires « générée-diffusée » ($\mathcal{H}_1$).

3.3 Classification multiclasse

Dans cette seconde expérience, nous nous intéressons à la classification multiclasse des images. Les mêmes combinaisons sont utilisées pour produire les résultats de la table 2. A nouveau, nous pouvons observer que l'agrégation des canaux couleurs des 3 générateurs « *image-to-image* » donne les meilleurs résultats. Dans ce cas, on observe que les résultats sont relativement similaires à ceux obetnus dans le cas binaire. Cela indique que les images générées par différents modèles sont discriminables entre elles autant que par rapport à des images réelles. La matrice de confusion du détecteur basé sur les 9 caractéristiques est donnée en figure 5. Elle montre que les images mal classées sont essentiellement issues de Stable-Diffusion 2.1 et de Pixart-α, les deux générateurs pour lesquels nous n'avons pas de modèle « *image-to-image* ». Nous pouvons donc émettre l'hypothèse qu'ajouter des modèles

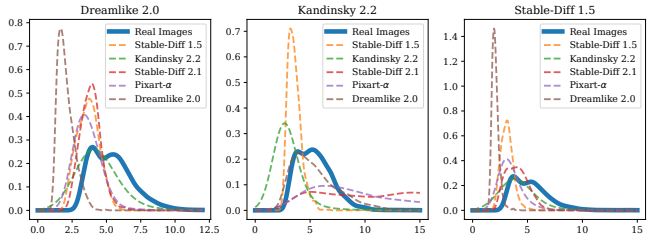

FIGURE 4 : Estimation de la densité de probabilité des distances géodésiques pour chaque générateur (en pointillés) et les images réelles (en bleu), pour la luminance des images obtenues avec les trois modèles « *image-to-image* ».

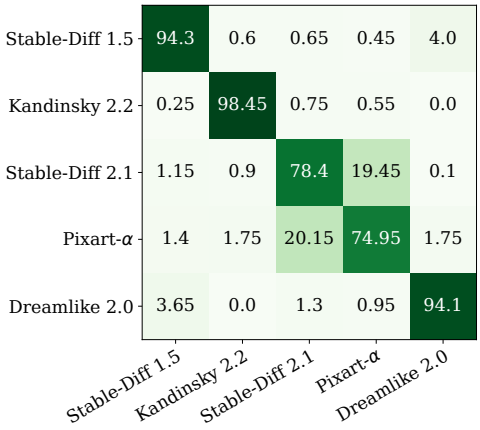

FIGURE 5 : Matrice de confusion du détecteur multiclasse basé sur les 9 caractéristiques.

« *image-to-image* » pour ces générateurs améliorerait les performances de la méthode. Notons enfin que la confusion entre Stable-Diffusion 1.5 et DreamLike-Photoreal 2.0 est relativement élevée, ce qui s'explique à nouveau par la similarité des deux modèles.

4 Conclusion et perspectives

Dans cette courte étude, nous avons proposé une méthode de détection d'images générées par des modèles de diffusion, basée sur la distance géodésique entre matrices de corrélation de résidus des images initiales et des images après une étape de diffusion. Nous avons montré que les distances calculées à partir de plusieurs modèles « *image-to-image* » et sur les différents canaux couleurs sont complémentaires pour la détection.

Les résultats obtenus sont encourageants, et laissent entrevoir des perspectives intéressantes. Notamment, en augmentant le nombre de modèles « *image-to-image* », mais aussi potentiellement en utilisant cette approche pour réaliser un apprentissage par contraste directement sur les matrices de corrélations.

Canal	Dreamlike	Kandinsky	Stable-Diff	Tous
Y	61.5	57.3	59.4	23.3
Cr	55.5	66.2	64.4	27.2
Cb	52.4	65.2	64.0	22.2
Tous	47.0	47.7	46.5	11.0

TABLE 2 : Probabilité d'erreur (en %) de la classification multiclasse pour chaque canal couleur et chaque générateur en entrée. Une validation croisée à 10 plis est utilisée.

Remerciements

Ce travail est réalisé grâce au financement du projet ANR PACeS No. ANR-21-CE39-0002, et du PEPR Cybersecurité Compromis France 2030.

Références

[1] Quentin BAMMEY : Synthbuster : Towards detection of diffusion model generated images. *IEEE OJSP*, 5, 2023.

[2] Alexandre BARACHANT, Stéphane BONNET, Marco CONGEDO et Christian JUTTEN : Multiclass brain–computer interface classification by riemannian geometry. *IEEE Trans. on Biomedical Engineering*, 2012.

[3] Lorenzo BARALDI, Federico COCCHI, Marcella CORNIA, Lorenzo BARALDI, Alessandro NICOLOSI et Rita CUCCHIARA : Contrasting deepfakes diffusion via contrastive learning and global-local similarities. *In European Conference on Computer Vision*. Springer, 2024.

[4] Rémi COGRANNE, Éva GIBOULOT et Patrick BAS : Alaska# 2 : Challenging academic research on steganalysis with realistic images. *In 2020 IEEE International Workshop on Information Forensics and Security (WIFS)*. IEEE, 2020.

[5] Riccardo CORVI, Davide COZZOLINO, Giovanni POGGI, Koki NAGANO et Luisa VERDOLIVA : Intriguing properties of synthetic images : from generative adversarial networks to diffusion models. *In Proc. of the IEEE/CVF conf. on CVPR*, 2023.

[6] Riccardo CORVI, Davide COZZOLINO, Giada ZINGARINI, Giovanni POGGI, Koki NAGANO et Luisa VERDOLIVA : On the detection of synthetic images generated by diffusion models. *In ICASSP 2023-2023 IEEE ICASSP*. IEEE, 2023.

[7] Davide COZZOLINO, Giovanni POGGI, Riccardo CORVI, Matthias NIESSNER et Luisa VERDOLIVA : Raising the bar of ai-generated image detection with clip. *In Proc. of the IEEE/CVF CVPR*, 2024.

[8] Quentin GIBOULOT : *Statistical Steganography based on a Sensor Noise Model using the Processing Pipeline*. Thèse de doctorat, Université de Technologie Troyes, 2022.

[9] Yanhao LI, Quentin BAMMEY, Marina GARDELLA, Tina NIKOUKHAH, Jean-Michel MOREL, Miguel COLOM et Rafael Grompone VON GIOI : Masksim : Detection of synthetic images by masked spectrum similarity analysis. *In Proc. of IEEE/CVF CVPR*, 2024.

[10] Antoine MALLET, Patrick BAS et Rémi COGRANNE : Statistical correlation as a forensic feature to mitigate the cover-source mismatch. *In Proc. of the 2024 ACM Workshop on Information Hiding and Multimedia Security*, 2024.

[11] Antoine MALLET, Minoru KURIBAYASHI, Rémi COGRANNE et Patrick BAS : AN ORIGINAL METHOD FOR DETECTION OF AI-GENERATED IMAGES BASED ON NOISE COVARIANCE. working paper or preprint, 2025.

[12] Francesco MARRA, Diego GRAGNANIELLO, Luisa VERDOLIVA et Giovanni POGGI : Do gans leave artificial fingerprints ? *In 2019 IEEE conf. on multimedia information processing and retrieval*. IEEE, 2019.

[13] Falko MATERN, Christian RIESS et Marc STAMMINGER : Exploiting visual artifacts to expose deepfakes and face manipulations. *In 2019 IEEE Winter Applications of Computer Vision Workshops*. IEEE, 2019.

[14] Lakshmanan NATARAJ, Tajuddin Manhar MOHAMMED, Shivkumar CHANDRASEKARAN, Arjuna FLENNER, Jawadul H BAPPY, Amit K ROY-CHOWDHURY et BS MANJUNATH : Detecting gan generated fake images using co-occurrence matrices. *arXiv*, 2019.

[15] Sheng-Yu WANG, Oliver WANG, Richard ZHANG, Andrew OWENS et Alexei A EFROS : Cnn-generated images are surprisingly easy to spot... for now. *In Proc. of the IEEE/CVF CVPR*, 2020.

[16] Zhendong WANG, Jianmin BAO, Wengang ZHOU, Weilun WANG, Hezhen HU, Hong CHEN et Houqiang LI : Dire for diffusion-generated image detection. *In Proc. of the IEEE/CVF ICCV*, 2023.