# Maximiser la marge pour une détection robuste des photomontages

Julien SIMON DE KERGUNIC[1] Rony ABECIDAN[1,2] Patrick BAS[1] Vincent ITIER[1,3]

[1]Centre de Recherche en Informatique, Signal et Automatique de Lille, Avenue Henri Poincaré, 59655 Villeneuve d'Ascq, France

[2]Label4.ai, 111 avenue Victor Hugo 75016 Paris

[3]IMT Nord Europe, Institut Mines-Télécom, Centre for Digital Systems, F-59000 Lille, France

**Résumé –** Malgré les progrès récents en détection de photomontages, les outils forensiques qui se basent sur l'apprentissage profond restent difficiles à exploiter en pratique, en raison de leur forte sensibilité aux données d'entraînement. Un simple post-traitement appliqué aux images d'évaluation peut suffire à dégrader leurs performances, compromettant leur fiabilité en contexte opérationnel. Dans cette étude, nous montrons qu'un même détecteur peut réagir différemment à des post-traitements inconnus selon les poids appris, malgré des performances similaires sur des données issues de la distribution d'entraînement. Ce phénomène s'explique par la variabilité des espaces latents induite par les entraînements, qui structurent différemment la séparation des classes. Nos expériences révèlent une corrélation marquée entre la distribution des marges latentes et la capacité de généralisation du détecteur. Nous proposons ainsi une méthode simple pour une détection de photomontages robuste à destination des praticiens : entraîner plusieurs variantes d'un même modèle et sélectionner celle maximisant la marge dans l'espace latent, afin d'accroître la robustesse face aux post-traitements.

**Abstract –** Despite recent progress in splicing detection, deep learning-based forensic tools remain difficult to deploy in practice due to their high sensitivity to training conditions. Even mild post-processing applied to evaluation images can significantly degrade detector performance, raising concerns about their reliability in operational contexts. In this work, we show that the same deep architecture can react very differently to unseen post-processing depending on the learned weights, despite achieving similar accuracy on in-distribution test data. This variability stems from differences in the latent spaces induced by training, which affect how samples are separated internally. Our experiments reveal a strong correlation between the distribution of latent margins and a detector's ability to generalize to post-processed images. Based on this observation, we propose a practical strategy for building more robust detectors: train several variants of the same model under different conditions, and select the one that maximizes latent margins.

## 1 Introduction

Un photomontage désigne toute modification structurée d'une image destinée à en altérer le sens, qu'il s'agisse de l'insertion d'éléments externes (*splicing*) ou de la duplication interne de régions (*copy-move* ou *cloning*). Dans la suite, nous nous concentrons sur le scénario *splicing*, cible privilégiée des détecteurs basés sur l'analyse du *bruit résiduel*. Les détecteurs actuels, majoritairement fondés sur des réseaux de neurones profonds, recherchent des incohérences *dans ce résidu* [1], [2]. Néanmoins, leurs performances en contexte réel se révèlent souvent inférieures à celles rapportées dans la littérature. Cet écart s'explique par l'application de post-traitements inconnus aux images falsifiées : une chaîne de développement suffit à induire un *décalage de domaine*, en modifiant la distribution statistique des images, authentiques comme falsifiées, et en dégradant la robustesse des détecteurs. Ces transformations (netteté, débruitage, etc.) altèrent le bruit et introduisent des dépendances entre zones originales et manipulées, rendant les photomontages moins détectables par les méthodes forensiques basées sur la détection d'anomalies locales. Par conséquent, tous les détecteurs de photomontages s'appuyant sur les résidus de bruits sont affectés par les post-traitements.

**Robustesse opérationnelle des détecteurs.** Dans la littérature forensique, le *décalage de domaine* induit par les post-traitements est un problème connu. Les approches *centrées sur les données* consistent à réfléchir à la construction d'une base d'entrainement pertinente pour maximiser la performance pratique des détecteurs [3]-[6], tandis que les approches *centrées sur le détecteur* visent à construire des détecteurs naturellement plus robustes face aux données hors distribution en utilisant par exemples des méthodes d'adaptation de domaines [7]-[9]. Ces méthodes reposent souvent sur des hypothèses fortes non validées en pratique comme l'équilibrage des classes en cible. À notre connaissance, la robustesse hors distribution sans accès aux cibles reste peu étudiée en détection de photomontage.

| Graine | Précision sur la source | Précision moyenne sur les cibles | Écart-type sur les cibles |
|---|---|---|---|
| 4 | 84% | 72% | 1.8 |
| 6 | 84% | 74% | 2.0 |
| 8 | 84% | 66% | 2.3 |

TABLE 1 : Impact de différentes graines d'initialisation sur les performances (hors distribution) du détecteur de Bayar [1]. La précision moyenne est obtenue en moyennant les performances d'un détecteur de Bayar, entraîné avec trois graines sur 20 cibles post-traitées via RawTherapee. Ensemble d'entraînement : $N_{\text{source}} \sim 20{,}000$ patches ; ensembles de test : $N_{\text{source}} \sim N_{\text{target}} \sim 7{,}000$ patches.

Le tableau 1 présente les performances d'un même détecteur de photomontage entraîné avec trois initialisations différentes. Bien que les performances soient similaires sur un ensemble d'images test suivant la même distribution que les images d'entraînement, les résultats varient grandement sur 20 versions post-traitées de ce même ensemble d'images. Cela montre que des entraînements différents qui conduisent à des performances équivalentes sur un ensemble de test suivant la

distribution d'entraînement, peuvent mener à des modèles inégalement robustes sur des échantillons hors distribution, du fait de la convergence vers des minimums locaux distincts. Cette disparité soulève des questions sur les pratiques d'entraînement à adopter pour renforcer la robustesse des détecteurs de photomontages face aux post-traitements.

**Contributions.** Nous cherchons à comprendre pourquoi une même architecture peut se montrer plus ou moins robuste aux post-traitements selon l'entraînement suivi. Dans un scénario réaliste, un expert peut entraîner plusieurs détecteurs et choisir le plus robuste, sans connaître la distribution d'évaluation. Pour cela, nous testons plusieurs entraînements d'un même modèle et analysons leurs comportements hors distribution.

1. Mise en évidence de l'effet négatif d'un surapprentissage à la source sur la généralisation du détecteur.
2. Découverte d'une corrélation claire entre la répartition des échantillons d'entraînement dans les espaces latents et la robustesse du détecteur face aux post-traitements.

La section 2 précise notre cadre d'étude ; la section 3 présente les expériences, et la section 4 discute des résultats et perspectives.

# 2 Formalisation

## 2.1 Formulation du problème et scénario

En adoptant les conventions de [10], une chaîne de traitement peut se représenter par un vecteur $\omega \in \Omega$ regroupant les paramètres d'une chaîne de développement (le coefficient de débruitage, le facteur de qualité JPEG, *etc.*). Pour la détection de photomontages, il est courant d'utiliser des modèles issus de l'apprentissage automatique :

$$f(x \mid \theta_\omega) : \mathcal{X} \to \{\text{authentique}, \text{falsifié}\}$$
$$x \mapsto y.$$

Ici, $\theta_\omega \in \Theta$ représente les paramètres appris à partir d'images authentiques et falsifiées ayant subi un post-traitement selon les paramètres $\omega$. Pour évaluer l'impact d'un décalage de domaine, il est courant de calculer l'*écart de généralisation* entre une source $s$ (base d'entraînement) et une cible $t$ (base d'évaluation) :

$$\mathcal{G}_{f(x|\theta_\omega)}(\omega_s, \omega_t) = \mathbb{E}_{(x,y)\sim P((x,y)|\omega_s)}\left(f(x \mid \theta_{\omega_s}) = y\right) - \mathbb{E}_{(x,y)\sim P((x,y)|\omega_t)}\left(f(x \mid \theta_{\omega_s}) = y\right). \quad (1)$$

Cet écart correspond à la différence de performance entre un scénario idéal — où source et cible partagent le même post-traitement — et un scénario réaliste où la distribution cible est inconnue. Dans notre cadre, les échantillons cibles ne sont pas accessibles à l'apprentissage ; l'objectif est donc de construire un détecteur de photomontages aussi robuste que possible à des post-traitements inconnus.

## 2.2 Marges dans les espaces latents

En détection de photomontage, nous considérons deux classes : *authentique* et *falsifiée*. En conséquence, nos modèles produisent deux scores de logit, $f_1$ et $f_2$, pour chaque entrée $x \in \mathcal{X}$. La classe prédite est celle ayant le score le plus élevé, c'est-à-dire $i^* = \arg\max_i f_i(x)$. Les détecteurs sont constitués de couches successives, chaque couche projetant son entrée dans un nouvel espace latent. La frontière de décision linéaire dans l'espace latent final apparaît non linéaire dans les espaces latents précédents, menant à une frontière de décision spécifique à chaque espace latent. La frontière de décision $\mathcal{D}^l$ du $l$-ième espace latent de notre détecteur est définie comme l'ensemble des points $x^l$ de cet espace pour lequel le détecteur est incertain entre les deux classes :

$$\mathcal{D}^l = \left\{x^l \mid f_1(x^l) = f_2(x^l).\right\} \quad (2)$$

On peut alors définir la marge d'un échantillon latent $x^l$ par rapport à cette frontière $\mathcal{D}^l$ comme étant la plus petite perturbation $\delta^l$ nécessaire pour amener $x^l$ sur la frontière de décision de l'espace latent $l$ :

$$d^p_{f,x^l} = \min_{\boldsymbol{\delta}} \|\delta^l\|_p \quad \text{t.q.} \quad f_1(x^l + \delta^l) = f_2(x^l + \delta^l). \quad (3)$$

L'écart de performance causé par les post-traitements résulte des biais spécifiques appris par les détecteurs de photomontages, dont les frontières de décision s'adaptent à la distribution source mais peinent à généraliser sur d'autres distributions. Intuitivement, des frontières trop proches des échantillons rendent le modèle plus facilement sensible aux variations induites par les post-traitements. Une étude précédente a justement montré une corrélation entre l'écart de généralisation et la distribution des marges latentes [11]. Toutefois, cette analyse se limitait à deux cibles et à des classifieurs d'images au sens sémantique.

# 3 Robustesse et marges latentes

## 3.1 Protocole expérimental

**Choix du détecteur et hyperparamètres.** Nos expériences s'appuient sur le détecteur de photomontages de Bayar et Stamm [1], un réseau de neurones convolutionnel largement utilisé par la communauté forensique. Son architecture, illustrée dans la figure 1, suit un schéma classique (Convolution + Max Pooling + Couches entièrement connectées), à l'exception de la première couche convolutionnelle, contrainte à effectuer un filtrage passe-haut :

$$\begin{cases} \boldsymbol{w}_k^{(1)}(0,0) = -1, \\ \sum_{m,n\neq 0} \boldsymbol{w}_k^{(1)}(m,n) = 1. \end{cases}$$

Cette contrainte favorise l'extraction de résidus de bruit trahissant des manipulations. Dans cet article, nous utilisons cette architecture pour analyser son comportement face à 20 cibles ayant subi des post-traitements variés. Les choix suivants ont été adoptés pour l'optimisation et les hyperparamètres :

- L'entraînement s'effectue sur un maximum de 115 époques, suffisant pour assurer la convergence.
- L'optimiseur utilisé est SGD.
- Le *batch size* est fixé à 128, un compromis adapté aux capacités d'un GPU classique et à la stabilité de l'apprentissage.
- Le *learning rate* initial est de $10^{-3}$, réduit d'un facteur 10 après 4 époques sans amélioration.

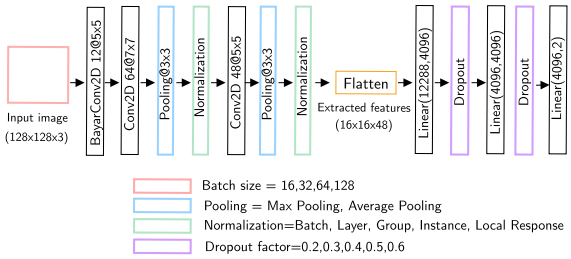


FIGURE 1 : Schéma des détecteurs de Bayar [1]. Les cellules colorées indiquent les hyperparamètres ou opérateurs modifiés au cours de nos 200 entraînements.

- Les poids sont initialisés selon les réglages par défaut de `PyTorch` [12].
- Toutes les expériences utilisent la graine 22 (choisie aléatoirement) pour assurer la reproductibilité et comparabilité des expériences.

Les ensembles source et cible sont divisés en sous-ensembles entraînement/validation/test selon la proportion $0.6/0.2/0.2$. Un mécanisme d'*early stopping*, basé sur la précision de l'ensemble de validation est utilisé pour éviter le sur-apprentissage.

**Construction des domaines source et cible.** Nous utilisons DEFACTO [13], le plus grand jeu de données public dédié à la détection de photomontages, en raison du réalisme des images et du contrôle offert sur les post-traitements via le format TIFF. La catégorie photomontage est divisée en deux sous-ensembles indépendants de taille équivalente : l'un pour la source, l'autre pour les cibles. Chaque image est découpée en patches de $128 \times 128$ pour un entraînement homogène. Un patch est étiqueté *falsifié* si la zone falsifiée couvre entre 10 % et 40 % de sa surface, une plage choisie pour éviter les cas extrêmes où le détecteur aurait du mal à distinguer les distributions de bruit. Des patchs *authentiques* sont ensuite ajoutés pour équilibrer les classes. Ce prétraitement génère environ 20 000 patches pour l'entraînement et 7 000 pour le test. La source est constituée uniquement d'images TIF originales, tandis que les cibles subissent divers post-traitements, appliqués directement sur les TIF avant découpage afin d'éviter les artefacts. Enfin, les cibles partagent un contenu visuel similaire, ce qui permet d'attribuer les écarts de performance uniquement aux effets des post-traitements, et non aux différences de contenu.

**Pipelines de post-traitement.** Nous définissons 20 chaînes de développement en combinant des opérations de débruitage par ondelettes et de renforcement de netteté via RawTherapee. Une compression JPEG avec un facteur de qualité de 70 est ensuite appliquée en fin de pipeline à l'aide de ImageMagick, assurant un contrôle complet sur le processus. Ce choix constitue un bon compromis entre réalisme et variabilité des cibles.

**Entraînements multiples d'une même architecture.** Nous proposons ici de faire varier les hyperparamètres et opérateurs du détecteur de Bayar afin d'étudier leur influence sur la capacité du modèle à généraliser à des échantillons post-traités. Nous réalisons 200 entraînements du détecteur de Bayar en modifiant la taille de batch, le type de pooling, la normalisation et le dropout, selon le schéma présenté en figure 1.

**Analyse par courbes de quantiles.** Les courbes de quantiles permettent de visualiser comment la distribution de l'écart de généralisation évolue dans une fenêtre glissante centrée sur des valeurs successives de diverses métriques. Pour chaque courbe de quantile, nous explorons les valeurs de la métrique avec un pas donné et une taille de fenêtre spécifiée (indiquée dans les légendes), en équilibrant le compromis entre localisation précise et fiabilité de l'estimation des quantiles.

## 3.2 Surapprentissage sur la source

Les 200 entraînements confirment le phénomène de surapprentissage à la source, évoqué dans [11] : plus l'accuracy sur l'ensemble de test source est élevée, plus l'écart de généralisation vers les cibles augmente. Ce comportement suggère que le réseau apprend progressivement des biais spécifiques à la source, optimisant ses performances au détriment de la robustesse sur les cibles.

## 3.3 Marges latentes et écarts de généralisation

Nous cherchons à vérifier s'il existe une corrélation entre les marges latentes des échantillons d'entrainement et les écarts de généralisation. Pour éviter les biais dus au sous-apprentissage, l'analyse est restreinte aux 138 modèles ayant atteint au moins 75 % de précision sur la source. Le calcul des marges latentes suit une méthodologie inspirée de [11] :

1. Estimer les marges latentes $d^2_{f,x^l}$ de chaque échantillon source $x^l$ à l'aide des logits $f_1$, $f_2$ et de leurs gradients par rapport à chaque couche, en prenant soin de les normaliser pour assurer l'indépendance à l'échelle. Nous excluons les marges négatives causées par les erreurs de classification. Davantage de détails sur ce calcul sont disponibles dans [11].
2. Résumer les distributions de marges pour chaque espace latent à l'aide de vecteurs $\boldsymbol{\mu_l}$ contenant des statistiques descriptives (premier et troisième quartiles, médiane, bornes supérieure et inférieure). Il est ensuite possible de concaténer tous les $\boldsymbol{\mu_l}$ dans un unique vecteur $\boldsymbol{\mu}$.
3. Combiner les statistiques de marge de chaque vecteur afin de dériver une métrique de marge $\mathcal{M}$. Nous suggérons de tester des combinaisons simples en calculant la somme des statistiques élevées à une certaine puissance $\alpha$ :
$$\mathcal{M}_\alpha = \sum_i \mu_i^\alpha. \quad (4)$$
L'élévation à la puissance $\alpha$ permet de mieux accentuer les différences de marge entre architectures.

Contrairement à [11], notre objectif n'est pas de prédire l'écart de généralisation à partir des marges, mais de fournir aux analystes forensiques une métrique pour évaluer la qualité de l'entraînement en vue d'une détection robuste de photomontage. Pour chaque détecteur de Bayar $f^i(x|\theta_s)$ entraîné sur notre source $s$ et chaque cible $t$ ayant subi un post-traitement via le pipeline $\omega_t$, nous générons des couples :
$$[\mathcal{M}_\alpha[f^i(x|\theta_s)]\ ,\ \mathcal{G}_{f^i(x|\theta_{\omega_s})}(\omega_s, \omega_t)]. \quad (5)$$
Les 2760 couples obtenus permettent d'étudier la corrélation entre $\mathcal{M}_\alpha$ et les écarts de généralisation. Les courbes de quantiles de la figure 2 montrent que les détecteurs les plus robustes

aux post-traitements sont ceux qui séparent le mieux leurs échantillons d'entraînement dans l'espace latent, en particulier avec $\mathcal{M}_2$.

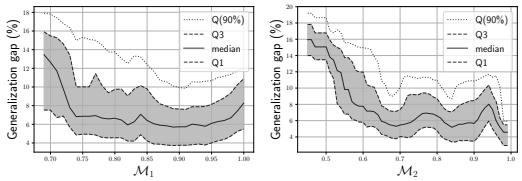


FIGURE 2 : Courbes de quantiles représentant l'évolution de l'écart de généralisation sur nos 20 domaines en fonction des métriques de marge $\mathcal{M}_1$ et $\mathcal{M}_2$, calculées à partir des marges latentes issues de toutes les couches. Ces métriques sont normalisées pour comparaison. Q1 est le premier quartile, Q3 est le troisième quartile et Q(90%) est le 90e percentile. Les points métriques sont explorés avec un pas de 0,01 et une taille de fenêtre de 0,1.

### 3.4 Importance de chaque marge latente

Bien que [11] souligne que l'analyse d'un seul espace latent soit insuffisante pour expliquer l'écart de généralisation, nous explorons ici la corrélation entre $\mathcal{G}$ et $\mathcal{M}_1$ en calculant les marges couche par couche. Ce choix est motivé par l'observation d'une corrélation positive entre ces deux métriques dans les cas de marges élevées, que nous cherchons à mieux comprendre.

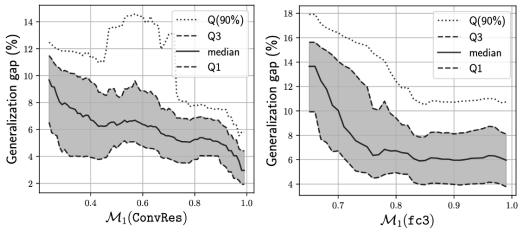


FIGURE 3 : Courbes de quantiles représentant l'évolution de l'écart de généralisation sur nos 20 domaines en fonction de $\mathcal{M}_1$, calculée à partir des marges latentes associées à la première couche (ConvRes) et dernière (fc3) de nos détecteurs de Bayar. Ces métriques sont normalisées pour comparaison. Q1 est le premier quartile, Q3 est le troisième quartile et Q(90%) est le 90e percentile. Les points métriques sont explorés avec un pas de 0,01 et une taille de fenêtre de 0,1.

La figure 3 montre une corrélation nette pour la première et la dernière couche latente. En revanche, nos expériences ne nous ont pas montré de corrélation dans les couches intermédiaires. Nous expliquons l'effet observé sur la première couche par le rôle des couches en amont, qui capturent des caractéristiques générales [14] : des marges élevées à ce niveau contribuent à la robustesse sur la cible. En sortie, la dernière couche étant la plus spécifique, de faibles marges y rendent le modèle sensible aux perturbations induites par les post-traitements. D'où l'intérêt d'un espace final bien séparateur, comme le suggèrent les approches basées sur les pertes contrastives [15]. L'absence de corrélation dans les couches intermédiaires s'explique sans doute par leur rôle transitoire : elles visent à préparer la séparation finale, sans structurer leur propre espace latent.

## 4 Conclusion

Cet article met en lumière la sensibilité des détecteurs de photomontage aux post-traitements inconnus, en montrant que des entraînements distincts d'une même architecture peuvent conduire à des capacités de généralisation très variables. Nos résultats soulignent qu'un surapprentissage à la source dégrade la robustesse hors distribution, justifiant l'utilisation de critères d'arrêt précoces adaptés. Nous proposons alors une métrique de marge latente, construite à partir de statistiques simples, et corrélée à l'écart de généralisation. Les marges issues des premières et dernière couches latentes sont particulièrement informatives. Nous recommandons ainsi d'entraîner plusieurs variantes d'un détecteur et de sélectionner celle maximisant les marges dans les couches clés. Dans nos travaux futurs, nous étendrons cette analyse à d'autres détecteurs et explorerons la conception d'architectures résilientes via des pertes contrastives, tout en étudiant le rôle des hyperparamètres dans la généralisation hors distribution.

## 5 Remerciements

Ces travaux ont bénéficié d'un accès aux moyens de calcul de l'IDRIS au travers de l'allocation de ressources 2025-AD011016555 attribuée par GENCI.

## Références

[1] B.BAYAR et M.C.STAMM, "A Deep Learning Approach to Universal Image Manipulation Detection Using a New Convolutional Layer", in *IH&MMSec '16*, 2016.

[2] F.GUILLARO, D.COZZOLINO, A.SUD, N.DUFOUR et L.VERDOLIVA, "TruFor : Leveraging all-round clues for trustworthy image forgery detection and localization", *IEEE CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[3] B.AHMED, T.A.GULLIVER et S.ALZAHIR, "Image splicing detection using mask-RCNN", *Signal, Image and Video Processing*, 2020.

[4] X.XU, J.DONG, W.WANG et T.TAN, "Robust steganalysis based on training set construction and ensemble classifiers weighting", in *2015 IEEE International Conference on Image Processing (ICIP)*, 2015.

[5] R.ABECIDAN, V.ITIER, J.BOULANGER, P.BAS et T.PEVNÝ, "Using Set Covering to Generate Databases for Holistic Steganalysis", in *IEEE International Workshop on Information Forensics and Security (WIFS 2022)*, 2022.

[6] R.ABECIDAN, V.ITIER, J.BOULANGER, P.BAS et T.PEVNÝ, "Leveraging Data Geometry to Mitigate CSM in Steganalysis", in *IEEE International Workshop on Information Forensics and Security (WIFS 2023)*, 2023.

[7] D.COZZOLINO, J.THIES, A.ROSSLER, C.RIESS, M.NIESSNER et L.VERDOLIVA, "ForensicTransfer : Weakly-supervised Domain Adaptation for Forgery Detection", *CoRR*, 2018.

[8] A.KUMAR et A.BHAVASAR, "Syn2Real : Forgery Classification via Unsupervised Domain Adaptation", *CoRR*, 2020.

[9] R.ABECIDAN, V.ITIER, J.BOULANGER et P.BAS, "Unsupervised JPEG Domain Adaptation for Practical Digital Image Forensics", in *IEEE International Workshop on Information Forensics and Security (WIFS 2021)*, 2021.

[10] D.ŠEPÁK, L.ADAM et T.PEVNÝ, "Formalizing cover-source mismatch as a robust optimization", in *EUSIPCO : European Signal Processing Conference*, 2022.

[11] Y.JIANG, D.KRISHNAN, H.MOBAHI et S.BENGIO, *Predicting the Generalization Gap in Deep Networks with Margin Distributions*, 2019.

[12] K.HE, X.ZHANG, S.REN et J.SUN, "Delving deep into rectifiers : Surpassing human-level performance on imagenet classification", in *Proceedings of the IEEE international conference on computer vision*, 2015.

[13] G.MAHFOUDI, B.TAJINI, F.RETRAINT, F.MORAIN-NICOLIER, J.L.DUGELAY et M.PIC, "DEFACTO : Image and Face Manipulation Dataset", in *27th European Signal Processing Conference (EUSIPCO 2019)*, 2019.

[14] J.YOSINSKI, J.CLUNE, Y.BENGIO et H.LIPSON, "How transferable are features in deep neural networks ?", *CoRR*, 2014.

[15] D.COZZOLINO, D.GRAGNANIELLO, G.POGGI et L.VERDOLIVA, "Towards Universal GAN Image Detection", 2021.