

## **INTITULÉ DU SUJET DE THÈSE**

Apprentissage profond pour la détection de falsifications et l'évaluation de la confidentialité dans les images numériques

## **ACRONYME DU PROJET**

MALEFICE - MACHine LEarning for Forensics and Image Confidentiality Evaluation

## **DIRECTEUR DE THÈSE**

Patrick Bas - DR CNRS, équipe SigMA (CRISTAL), [patrick.bas@cnrs.fr](mailto:patrick.bas@cnrs.fr)

## **ENCADRANT DE THÈSE**

Pauline Puteaux - CR CNRS, équipe SigMA (CRISTAL), [pauline.puteaux@cnrs.fr](mailto:pauline.puteaux@cnrs.fr)

## **UNITÉ DE RECHERCHE**

CRISTAL, UMR CNRS 9189  
Bâtiment ESPRIT  
Avenue Henri Poincaré  
59655 Villeneuve d'Ascq, France

## **RÉSUMÉ DE LA THÈSE**

Dans cette thèse, nous proposons d'utiliser l'apprentissage automatique dans le contexte de la sécurité multimédia, et plus particulièrement pour analyser les images qui sont chiffrées sélectivement (seules certaines composantes des images sont chiffrées afin de préserver leur format et/ou partiellement leur sémantique). Les méthodes d'apprentissage automatique sont devenues en quelques années l'état de l'art en matière de détection de falsifications et de reconnaissance de formes. Elles sont également utilisées pour évaluer la quantité d'information qui peut fuir dans un système de chiffrement. Nous nous intéressons au compromis entre la détection de falsifications dans le domaine chiffré sélectivement et la protection de la vie privée (en termes de confidentialité).

## **MOTS-CLÉS**

Apprentissage profond ; analyse et traitement des images dans le domaine chiffré ; détection de falsifications ; sécurité des données multimédia ; images numériques ; protection de la vie privée.

## **CONTEXTE DE LA RECHERCHE**

Ce projet s'inscrit dans une démarche cohérente dans l'équipe SigMA autour du transfert des approches de l'apprentissage automatique au domaine de la sécurité. Le sujet de thèse présenté propose d'adapter les approches de l'apprentissage automatique au domaine de la sécurité. Ce projet peut contribuer à développer l'intelligence artificielle dans la Région Hauts de France et répond aux ambitions du rapport Villani (« aiforhumanity »). Par ailleurs, il s'inscrit dans les efforts nationaux autour de la cybersécurité. Cette dynamique se concrétise avec le lancement, à l'initiative du chef de l'état, du Campus Cyber, dont le laboratoire CRISTAL est membre. Par ailleurs, il s'inscrit dans les thématiques du hub HumAIn@Lille duquel Patrick Bas a participé au renforcement via le dépôt d'une chaire. En outre, il s'inscrit dans les axes 1 et 2 définis dans la chaîne de valeur du CPER Cornelia :

- Axe 1 : Bases théoriques et scientifiques de l'IA - Sécurité et Robustesse aux attaques
- Axe 2 : L'intelligence artificielle embarquée et enjeux sociétaux - Sécurité des données en environnement physique contraint.

Ce projet pourra apporter des solutions viables pour assurer une « transmission de données fiables et non falsifiables », enjeu tout particulièrement mis en avant dans l'axe 2. Il permettra également de participer à l'axe 1 en proposant de détecter les falsifications d'images.

## PRÉSENTATION RAPIDE DES ENCADRANTS

Patrick Bas est Directeur de Recherche CNRS au laboratoire CRISAL. Ses recherches portent sur la sécurité des contenus avec comme applications la stéganographie (insertion d'informations cachées) et la détection de signaux faibles comme la stéganalyse (détection d'informations cachées) ou l'analyse forensique. Patrick Bas a proposé une chaire IA en 2019.

Pauline Puteaux travaille en tant que chargée de recherche CNRS au CRISAL, dans l'équipe SigMA (depuis novembre 2021). Ses recherches portent sur la sécurité multimédia, et plus particulièrement sur l'analyse, le traitement et la protection des données multimédia dans le domaine chiffré (insertion de données cachées dans des images chiffrées, correction d'images chiffrées bruitées, recompression d'images crypto-compressées, détection de falsifications dans le domaine chiffré...). Elle a obtenu son doctorat en informatique à l'Université de Montpellier en 2020.

## PRÉSENTATION DU PROJET, DE LA MÉTHODOLOGIE, D'UN ENSEMBLE DE RÉFÉRENCES

Dans cette thèse, nous proposons d'utiliser l'apprentissage automatique dans le contexte de la sécurité multimédia, et plus particulièrement pour analyser les images qui sont chiffrées sélectivement (seules certaines composantes des images sont chiffrées afin de préserver leur format et/ou partiellement leur sémantique). Les méthodes d'apprentissage automatique sont devenues en quelques années l'état de l'art en matière de détection de falsifications et de reconnaissance de formes. Elles sont également utilisées pour évaluer la quantité d'information qui peut fuir dans un système de chiffrement. Nous nous intéressons au compromis entre la détection de falsifications dans le domaine chiffré sélectivement et la protection de la vie privée (en termes de confidentialité) [Puteaux2021].

Les échanges d'images représentent aujourd'hui une part importante de l'utilisation d'Internet. Cette tendance va de pair avec des exigences de confidentialité puisque la transmission peut être espionnée sur les canaux publics. Dans ce contexte, il a été proposé de chiffrer les images afin de dissimuler leur contenu et de les rendre visuellement confidentielles pour les utilisateurs non-autorisés.

Certaines méthodes de chiffrement ont été spécifiquement conçues pour les images afin de préserver leur format et leur taille et de permettre leur visualisation après chiffrement. En particulier, le chiffrement sélectif, qui ne protège qu'une partie des informations de l'image, permet de visualiser un niveau de détails de l'image en fonction de la quantité d'informations chiffrées [VanDroogenbroeck2002].

Pour les utilisateurs finaux tels que les plateformes en nuage (*cloud*) ou les réseaux sociaux, les images chiffrées ne sont pas faciles à analyser. En effet, en utilisant un schéma de chiffrement classique, la plateforme ciblée n'est pas en mesure de décider si une image respecte ou non ses conditions d'utilisation. En particulier, elle ne peut pas s'assurer que l'image n'a pas été falsifiée, c'est-à-dire déterminer si elle a été altérée ou non par une modification locale visant à changer la sémantique de l'image. Les falsifications peuvent être de nature endogène (copier/déplacer à partir d'une seule image), de nature exogène (copier/coller entre plusieurs images) ou de la synthèse d'images (inpainting, deepfakes). Les méthodes obtenant les meilleures performances pour la détection, la localisation et la caractérisation du type de falsification s'appuient sur des réseaux de neurones convolutionnels (CNN) [Wu2019]. Si, dans le domaine clair, l'intégralité du contenu de l'image est considéré lors de cette analyse, cette stratégie ne peut pas être utilisée dans le domaine chiffré. En effet, le chiffrement ajoute un bruit pseudo-aléatoire de forte amplitude qui rend complexe l'extraction de caractéristiques significatives pour une classification comme authentique ou falsifiée.

Afin de préserver la confidentialité tout en permettant l'analyse dans le domaine chiffré, le chiffrement homomorphe peut être utilisé [Paillier1999]. Par exemple, la détection de points d'intérêt dans les images chiffrées, tels que les points SIFT, a été proposée [Hsu2011]. Cependant, les algorithmes de chiffrement homomorphes sont très lents, nécessitent beaucoup de ressources et augmentent la taille des données chiffrées de façon significative. Il est donc

difficile d'effectuer des opérations complexes - en particulier, d'appliquer des approches par apprentissage automatique - dans le domaine homomorphe.

A l'inverse, le chiffrement sélectif est rapide et permet de préserver la taille et le format de l'image originale. Avec une telle approche, une partie du contenu de l'image est chiffrée, tandis que l'autre reste en clair, c'est-à-dire non chiffrée, et peut ensuite être analysée. Si ces propriétés sont intéressantes, elles mettent en évidence une potentielle faille de sécurité : la confidentialité visuelle du contenu de l'image peut être menacée.

Dans la littérature, l'évaluation du niveau de confidentialité visuelle s'effectue généralement par le biais d'une analyse statistique (calcul de corrélations, d'entropie de Shannon, test du Chi-2, évaluation de la qualité visuelle avec référence à l'image originale [Wu2011]). Cependant, il a été montré qu'obtenir de bons résultats à ces tests statistiques et à ceux mis en place par le NIST [Rukhin2001] est une condition nécessaire mais non suffisante pour prouver qu'une méthode de chiffrement d'images est sécurisée [Preishuber2018]. En effet, ils ne tiennent pas compte des scénarios de sécurité où les attaquants utilisent la connaissance de l'algorithme de chiffrement lors de leur attaque. De plus, ils ne permettent pas d'évaluer la « reconnaissabilité » du contenu, c'est-à-dire la capacité à détecter et à localiser des objets ou des personnes dans une image chiffrée sélectivement.

Par ailleurs, dans le but de s'inscrire dans un cadre opérationnel, il sera important de considérer les formats d'échange et de stockage d'images les plus classiques, comme le format de compression JPEG [Wallace1992] sur lequel reposent plusieurs méthodes de chiffrement sélectif [Rodrigues2006, Li2007, Unterweger2012]. Les spécificités de ce format doivent être prises en compte lors de la détection de falsifications (il peut être source d'information par rapport aux manipulations [Bianchi2012, Iakovidou2018]), mais aussi lors de l'évaluation de la confidentialité visuelle.

L'étudiant(e) recruté(e) devra posséder des connaissances et des compétences solides en apprentissage automatique et être sensibilisé aux problématiques et aux enjeux de la sécurité des données. Une formation en traitement des images et en théorie de l'information sera un plus indéniable. Par ailleurs, une maîtrise du langage de programmation Python et, en particulier, de TensorFlow/Pytorch sera demandée.

## **INDICATIONS SUR LE DÉROULEMENT PRÉVU DE LA THÈSE**

Le déroulement de la thèse suivra le programme de travail suivant :

- T0-T0+6 : Etude bibliographique (détection de falsifications avec des réseaux de neurones convolutionnels, chiffrement sélectif d'images,)
- T0+6-T0+12 : Adaptation de méthodes de détection de falsifications basées apprentissage automatique à une utilisation dans le domaine chiffré sélectivement :
  - Identifier si une image est authentique ou falsifiée
  - Déterminer l'emplacement de la zone falsifiée
  - Déterminer le type de falsification
- T0+9-T0+15 : Recherche et mise en place d'un protocole d'évaluation du niveau de confidentialité visuelle d'une image chiffrée sélectivement :
  - Prise en compte des scénarios d'attaque, en particulier impliquant l'utilisation de réseaux antagonistes génératifs (GAN)
  - Évaluation de la « reconnaissabilité » par réseaux de neurones convolutionnels
  - Adaptation des méthodes statistiques existantes • Mise en place de tests d'hypothèse adaptés
- T0+15-T0+24 : Adaptation des méthodes développées au format d'images compressées JPEG
- T0+12-T0+32 : Valorisation (rédaction d'articles, participations à des conférences, vulgarisation)

L'étudiant(e) recruté(e) sera accueilli(e) au laboratoire CRISAL. Il/elle sera incité(e) à postuler à des bourses de séjour pour rejoindre des équipes internationales possédant une expertise forte dans le domaine, comme l'équipe d'Andreas Uhl (Salzburg University, Autriche).

## RÉFÉRENCES BIBLIOGRAPHIQUES

- [Bianchi2012] T. Bianchi and A. Piva. (2012). Image forgery localization via block-grained analysis of JPEG artifacts. *IEEE Transactions on Information Forensics and Security*, 7(3): 1003-1017.
- [Hsu2011] C.-Y. Hsu, C.-S. Lu, and S.-C. Pei, « Homomorphic encryption-based secure SIFT for privacy-preserving feature extraction, » in *Media Watermarking, Security, and Forensics*, vol. 7880. International Society for Optics and Photonics, 2011, p. 788005.
- [Iakovidou2018] C. Iakovidou, M. Zampoglou, S. Papadopoulos and Y. Kompatsiaris. (2018). Content-aware detection of JPEG grid inconsistencies for intuitive image forensics. *Journal of Visual Communication and Image Representation*, 54: 155-170.
- [Li2007] W. Li and Y. Yuan, « A leak and its remedy in JPEG image encryption, » *International Journal of Computer Mathematics*, vol. 84, no. 9, pp. 1367–1378, 2007.
- [Paillier1999] P. Paillier, « Public-key cryptosystems based on composite degree residuosity classes, » in *International Conference on the Theory and Applications of Cryptographic Techniques (EUROCRYPT)*. Springer 1999, pp. 223–238.
- [Preishuber2018] M. Preishuber, T. Hütter, S. Katzenbeisser, and A. Uhl (2018). Depreciating motivation and empirical security analysis of chaos-based image and video encryption. *IEEE Transactions on Information Forensics and Security*, 13(9): 2137–2150.
- [Puteaux2021] P. Puteaux, V. Itier, and P. Bas. (2021). Combining Forensics and Privacy Requirements for Digital Images. *arXiv preprint arXiv:2103.03569*.
- [Rodrigues2006] J. M. Rodrigues, W. Puech, and A. G. Bors, « Selective encryption of human skin in JPEG images, » in *International Conference on Image Processing (ICIP)*. IEEE, 2006, pp. 1981–1984.
- [Rukhin2001] A. Rukhin, J. Soto, J. Nechvatal, M. Smid, and E. Barker (2001). A statistical test suite for random and pseudorandom number generators for cryptographic applications. Technical report, Gaithersburg, MD, USA.
- [Unterweger2012] A. Unterweger and A. Uhl, “Length-preserving Bit-stream-based JPEG Encryption,” in *Proceedings of the on Multimedia and security*. ACM, 2012, pp. 85–90.
- [VanDroogenbroeck2002] M. Van Droogenbroeck and R. Benedett, « Techniques for a selective encryption of uncompressed and compressed images, » in *International Conference on Advanced Concepts for Intelligent Vision Systems (ACIVS)*, 2002, pp. 90–97.
- [Wallace1992] G. K. Wallace. (1992). The JPEG still picture compression standard. *IEEE Transactions on Consumer Electronics*, 38(1): XVIII-XXXIV.
- [Wu2011] Y. Wu, J. P. Noonan, and S. Aghaian (2011). NPCR and UACI randomness tests for image encryption. *Cyber journals: Multidisciplinary Journals in Science and Technology, Journal of Selected Areas in Telecommunications (JSAT)*, 1(2):31–38.
- [Wu2019] Y. Wu, W. AbdAlmageed, and P. Natarajan, “Mantra-net: Manipulation tracing network for detection and localization of image forgeries with anomalous features,” in *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2019, pp. 9543–9552.